A Genetical Genomics Approach to Genome Scans Increases Power for QTL Mapping

Guoying Sun and Paul Schliekelman¹

Department of Statistics, University of Georgia, Athens, Georgia 30602

Manuscript received October 8, 2010

Accepted for publication December 3, 2010

ABSTRACT

We describe a method for integrating gene expression information into genome scans and show that this can substantially increase the statistical power of QTL mapping. The method has three stages. First, standard clustering methods identify small (size 5–20) groups of genes with similar expression patterns. Second, each gene group is tested for a causative genetic locus shared with the clinical trait of interest. This is done using an EM algorithm approach that treats genotype at the putative causative locus as an unobserved variable and combines expression information from all of the genes in the group to infer genotype information at the locus. Finally, expression QTL (eQTL) are mapped for each gene group that shares a causative locus with the clinical trait. Such eQTL are candidates for the causative locus. Simulation results show that this method has far superior power to standard QTL mapping techniques in many circumstances. We applied this method to existing data on mouse obesity. Our method identified 27 putative body weight QTL, whereas standard QTL mapping produced only one. Furthermore, most gene groups with body weight QTL included *cis* genes, so candidate genes could be immediately identified. Eleven body weight QTL produced 16 candidate genes that have been previously associated with body weight or body weight-related traits, thus validating our method. In addition, 15 of the 16 other loci produced 32 candidate genes that have not been associated with body weight. Thus, this method shows great promise for finding new causative loci for complex traits.

IN the past few years, there has been a great deal of interest in the relationship between genotypic variation and transcriptional variation. Numerous studies have found that there is often a strong genotypic component to variation in transcription (*e.g.*, BREM *et al.* 2002; OLEKSIK *et al.* 2002; YAN *et al.* 2002; CHEUNG *et al.* 2003; SCHADT *et al.* 2003; MORLEY *et al.* 2004). These observations have spurred interest in the possibility of leveraging this strong relationship between genotype and transcription to explore the relationship between genotype and a clinical trait of interest, the premise being that genome-wide expression data provide access to the gene network that links genotype to clinical phenotype (for simplicity, we refer to the organismal trait of interest as the “clinical trait,” to distinguish it from an expression-level trait; of course, it need not be solely of clinical interest). Numerous studies have successfully used such approaches to find causative polymorphisms for complex traits (*e.g.*, LAWN *et al.* 1999; BERGE *et al.* 2000; KARP *et al.* 2000; WAYNE and MCINTYRE 2002; KIRST *et al.* 2004; BYSTRYKH *et al.* 2005; CHESLER *et al.* 2005; HUBNER *et al.* 2005; LI *et al.* 2005;

SCHADT *et al.* 2005, 2008; GHAZALPOUR *et al.* 2006; SCHADT 2006; CHEN *et al.* 2008; EMILSSON *et al.* 2008), showing the power of this type of approach.

To date, most work in this area has been based on prior knowledge of a genomic region containing a QTL for the trait of interest. The most common approaches are to look for genes in the region of interest whose expression levels are correlated with the clinical trait of interest or to look for gene networks with many expression QTL (eQTL) in the region of interest. Our focus in this article is the traditional genome-scan setting—that is, gene mapping with no prior knowledge of any QTL position. There are numerous genetical genomics articles (*e.g.*, GHAZALPOUR *et al.* 2005, 2006; AYROLES *et al.* 2009; EDWARDS *et al.* 2009) that are essentially genome scans in the sense of identifying QTL locations without using any prior QTL information. However, there is little understanding of how these approaches compare to traditional QTL mapping, what conditions (if any) make them superior, and what methods for using expression information are best and why. The goals of this article are to (1) develop a self-contained method for integrating expression information into genome scans, (2) explore the assumptions and the statistical properties of this method, and (3) show that this approach can substantially increase the power for QTL mapping.

Supporting information is available online at <http://www.genetics.org/cgi/content/full/genetics.110.123968/DC1>.

¹Corresponding author: Department of Statistics, 203 Statistics Bldg., University of Georgia, Athens, GA 30602-7952.
E-mail: pdschlie@stat.uga.edu

FIGURE 1.—Hypothetical schematic of a genetic pathway for a clinical trait. Ovals represent causative loci, that is, loci with genotype variation that leads to variation in the trait. Rectangles represent genes in a network, whose transcript variation between sample individuals leads to variation in clinical trait values between individuals. Arrows between loci and genes in the network represent an impact of genotypic variation at the locus on the gene's transcript level (note that some genes could be the loci themselves). Arrows between genes in the network represent control of transcription. Arrows between genes and the clinical trait represent an impact of the gene's transcription on the trait value.

Figure 1 shows a hypothetical schematic of the relationship between a clinical trait, a gene network, and the causative loci for the clinical trait. Genetic variation at the causative loci leads to changes in transcription in the gene network. These in turn lead to variation in the clinical trait. In traditional gene mapping, we have genotype data for markers (some of which are hopefully in linkage or association with a clinical trait locus) and we have clinical trait data. The main obstacle is that for a complex trait, the statistical link is weak between a causative locus and the trait because individual loci account for only a small part of the variation. The statistical link is further weakened by the nonperfect correlation between marker and causative locus.

Now, suppose that we also have information about the gene network, in the form of gene expression information. The network provides a link between the clinical trait and the causative loci. Consider the circled gene in Figure 1. The transcript level of this gene is affected by genotype variation at three of the causative loci and is thus a three-locus quantitative trait. Because it shares three controlling loci with the clinical trait, its genotype-based variation is in higher concordance with the trait than any single locus can be. SCHLIEKELMAN (2008) modeled this situation for a binary disease and showed that such a gene will have substantially higher association with the disease than will the causative loci. This is the fundamental idea of genetical genomics: that is, that the genes in the network provide an intermediate link between clinical trait and genotype and genome-wide

expression data can be used to access their transcript levels.

SCHLIEKELMAN (2008) showed that there are substantial statistical obstacles to this idea. As we move through the network farther away from the causative loci and closer to the trait, the statistical link to the trait becomes stronger and the link to the loci becomes weaker. We must establish a statistical link in both directions, but there is a trade-off in statistical power between these two goals. The required sample sizes for reasonable power for both directions using currently available methods range from hundreds to thousands, depending on assumptions about the genetics of transcription and disease.

In this article we present a new method for establishing a link between a gene transcript and a clinical trait that greatly increases statistical power relative to the results of SCHLIEKELMAN (2008). This article addresses quantitative clinical traits, while a future article will address binary ones. This method infers the genotype of the controlling locus for a group of correlated transcripts and tests whether that genotype is associated with trait status. We model the expression level of a gene as a normal mixture model with components corresponding to genotype. That is, the expression level has a normal distribution with mean and variance dependent on genotypes at the controlling loci for that transcript. A number of other authors have previously modeled expression levels this way in the eQTL context (*e.g.*, JIA and XU 2005; KENDZIORSKI *et al.* 2006; NETTLETON and WANG 2006; CHEN and KENDZIORSKI 2007; GELFOND *et al.* 2007; PEREZ-ENCISO *et al.* 2007; SAMPSON and SELF 2008; SILLANPAA and NOYKOVA 2008). HSIEH *et al.* (2007) found evidence of multimodality in transcription in 7–10% of genes in two *Drosophila melanogaster* lines. Other genes may have had either little within-population variation or genetic control too complicated to be detected with the available sample sizes.

We define a controlling locus for a transcript as a locus with genotypic variation between individuals that contributes to variation in the transcript between individuals. If there were sufficient separation between the components of the mixture, then we could use the expression level to assign an individual's genotype at the controlling locus with high probability. Consider Figure 2. This shows a scatter plot of a clinical trait value *vs.* transcript value for a simulated data set. The clinical trait and the transcript are jointly controlled by a locus D with alleles *D* and *d*. The expression means are 0, 0.5, and 1 for the *dd*, *Dd*, and *DD* genotypes, respectively. The standard deviation is 0.1 for each component and is small relative to the differences between components. Thus, given this plot, we could infer genotypes at the D locus for most sample individuals with high certainty. In this case, we could test whether the locus D that controls the transcript is a QTL for the clinical trait by assigning individuals to genotype groups according to

FIGURE 2.—Scatter plot of a clinical trait value *vs.* transcript value for a simulated data set. The clinical trait and the transcript are jointly controlled by a locus D with alleles *D* and *d* at frequencies of 0.5 each. The trait and expression means are both 0, 0.5, and 1 for the *dd*, *Dd*, and *DD* genotypes, respectively. The standard deviation is 0.1 for each component. The sample size was 200. Genotypes were sampled for each individual on the basis of the allele frequencies assuming Hardy–Weinberg equilibrium. Expression values and trait values were generated from a multivariate normal with means determined by genotype.

their expression values and then testing for an association between the assigned genotypes and the clinical trait values. We did this for the simulated data in Figure 2 with a sample size of 100, using the *k*-means algorithm to assign genotype and an ANOVA to test for association. There was a highly significant association between assigned genotype and the clinical trait, with a *P*-value of 2×10^{-8} .

This is, in effect, what the hypothesis test discussed in SCHLIEKELMAN (2008) is doing. Unfortunately, there is usually not this level of separation between components and a single expression level is not a reliable indicator of underlying genotype. In this article we present a method that combines information from multiple transcripts to infer the controlling genotype more reliably. Figure 3a shows a scatter plot for a simulated data set like that in Figure 2. In this example, the expression standard deviation is 0.5. Thus, there is much less separation between components and the individual genotypes cannot be distinguished. Visually, we can discern only a weak relationship between the transcript and the trait. When we apply the *k*-means algorithm to assign genotypes and an ANOVA to test for association with the clinical trait, we get a *P*-value for association of 0.39 with a sample size of 200. However, in Figure 3b we include a second transcript also controlled by the locus D. The *x*-axis and *y*-axis show values of the two transcripts

and the size of the point shows the value of the clinical trait for that individual. We can see a clear increasing pattern in the clinical trait value as we move along the diagonal line from low (*x*, *y*) values to high ones. That is, the joint distribution of the transcripts has a much stronger relationship with genotype than does a single transcript and therefore has a stronger relationship with the clinical trait.

In effect, each expression value is an independent observation of the underlying genotype. Because of the substantial variation in expression no single such observation is very reliable. However, by combining observations of multiple transcripts controlled by the same locus, we can get more reliable inference of the underlying genotype. The values *C1*, *C2*, and *C3* show the cluster centers when the *k*-means clustering algorithm was applied to the data with three clusters specified. We again used the cluster assignments as a proxy for genotype and tested for an association between inferred genotype and the clinical trait. The result was highly significant (*P*-value = 4×10^{-5} with sample size of 200).

In this article we formalize this idea using an approach based on the EM algorithm. The goal is a method for conducting genome scans. We assume that the data consist of (1) genotype data for a large number of markers spaced throughout the genome for each sample individual, (2) expression data from a microarray covering thousands of genes for each individual, and (3) the clinical trait value for each individual. We undertake a three-part procedure:

1. We apply hierarchical clustering to the expression data and apply an algorithm that extracts gene groups of a target size *w* from the bottom of the dendrogram. This produces groups of size *w* or smaller with high intergroup correlation.
2. We assume that genes in these groups share some common controlling factor. We then test whether this controlling factor affects the clinical trait. This factor could be environmental or genetic, but our method assumes that it is a genetic locus. Our approach uses the EM algorithm and treats genotype at the unobserved locus as the unobserved data and this is used as the basis of a likelihood-ratio test.
3. If the controlling factor for a gene group is determined to be causative for the clinical trait, then there are two options. One is to map eQTL for the gene group. This approach is pursued in this article. Any such eQTL will then become candidate genes for the trait. Another option is to then use the expression data to infer the genotype at the putative controlling locus and then use this genotype information to construct an optimal sample for mapping the locus, similarly to the informal procedure used by SCHADT *et al.* (2003). This idea will be developed in a future publication (E. H. BREAZEL and P. SCHLIEKELMAN, unpublished data)].

FIGURE 3.—(A) Scatter plot identical to Figure 2 except that the standard deviation is increased to 0.54. B includes a second transcript also controlled by the locus D. The x -axis and y -axis show values of the two transcripts and the size of the point shows the value of the clinical trait for that individual. The values C1, C2, and C3 show the cluster centers when the k -means clustering algorithm was applied to the data with three clusters specified.

There are several advantages to using expression data to infer genotype, rather than using marker data. The first is that by using expression data we are inferring the genotype of the actual locus and not just a marker linked to it. That is, there is no recombination between marker and locus to reduce power. Of course, we then have error resulting from variation in expression. In this sense, our method trades variation in recombination between marker and genotype for variation in expression. We will show evidence that the genotype can be inferred more accurately using expression information, at least in some circumstances. The second advantage is that, to the extent that gene expression is a good indicator of genotype, then we are directly testing association of genotype to the clinical trait. It is well established that association tests are superior to linkage tests for detecting common variants (RISCH and MERIKANGAS 1996). However, association tests rapidly lose power when there are multiple alleles at the locus (SLAGER *et al.* 2000). Our method collapses all alleles into either expression-increasing or expression-decreasing alleles. Thus, it is similar to an association test with two alleles, regardless of how many alleles there really are. The third advantage (discussed in SCHLIEKELMAN 2008) is that transcripts that share multiple controlling loci with the clinical trait have a higher correlation with the trait than do the loci themselves. In this case, the expression level is indicating multilocus genotype and a multilocus genotype has stronger association with the clinical trait than a single-locus genotype. These points are further addressed in the DISCUSSION.

The remainder of this article is organized as follows: First, we describe the method. Then, we show simulation results that explore the statistical power of the method. Finally, we apply the method to a mouse obesity data set. This is followed by a DISCUSSION section.

METHODS

Likelihood-ratio test: We assume that there are N individuals with a quantitative clinical trait that is controlled by genotype at L causative loci. The expression levels of s genes have been measured for each individual. Denote X_{ij} as the expressional level of the j th gene in the i th individual, and Y_i as the clinical trait value of the i th individual. Both X_{ij} and Y_i are treated as continuous random variables.

Our purpose is to test whether a controlling locus for a group of highly correlated genes is also a causative locus for a clinical trait. We assume that the n transcripts in the group share at least one controlling locus. If there is some environmental factor that affects the clinical trait and affects expression of the transcript, our method cannot distinguish this from a genetic factor with the same effect.

The hypothesis test of interest is H_0 , the controlling locus for the group of n transcripts is not a QTL for the clinical trait, and H_1 , the controlling locus for the group of n transcripts is also a QTL for the clinical trait.

The likelihood function under H_1 is

$$L_1(\Theta | \vec{X}, \vec{Y}) = \prod_{i=1}^N \sum_{g_i} f(\vec{X}_i | g_i) f(Y_i | g_i) P(g_i).$$

The likelihood function under H_0 is

$$L_0(\Theta | \vec{X}, \vec{Y}) = \prod_{i=1}^N \sum_{g_i} f(\vec{X}_i | g_i) f(Y_i) P(g_i).$$

$\vec{X}_i = (X_{i1}, X_{i2}, \dots, X_{in})^T$ is the vector of expression levels of the n genes in the group for the i th individual. The distribution $f(\vec{X}_i | g_i)$ is the expression-level probability density function (pdf) of the gene group. This is assumed to be a multivariate normal pdf whose mean and variance are dependent on the genotype g_i at the controlling locus. The clinical trait distribution is also assumed to be normal and depends on the genotype g_i at the controlling locus ($f(Y_i | g_i)$) under the alternate hypothesis, but not under the null ($f(Y_i)$). The test statistic is the ratio of the maximum likelihood under H_1 and under H_0 .

We evaluate evidence for association between transcripts and the clinical trait with the likelihood-ratio statistic $2[\text{Log}(\hat{L}_1(\Theta | \vec{X}, \vec{Y})) - \text{Log}(\hat{L}_0(\Theta | \vec{X}, \vec{Y}))]$. Rejecting the null hypothesis suggests that a controlling locus for these transcripts is also a QTL for the clinical trait.

EM algorithm: We use the EM algorithm (Dempster *et al.* 1977) to obtain maximum-likelihood estimates (MLE) of parameters since we do not actually observe the genotype. Matrix X and vector Y are treated as observed data. Genotype $\vec{G} = (g_1, g_2, \dots, g_N)^T$ is the unobserved data, where $g_i = 0, 1, 2$, is the number of transcript-increasing alleles in individual i at the controlling locus. In the results shown in this article, we assume that the expression-level distribution and the clinical trait distribution are independent after conditioning on the genotype at the controlling locus, that expression levels are independent within and between individuals after conditioning on genotype, and that the clinical trait is independent between individuals. In this case, the conditional expression and clinical trait distributions simplify to a product of independent normal pdfs. However, the first two of these assumptions are likely not true in many cases. That is, there will be correlations beyond shared genotypes between many transcripts and between transcripts and the clinical trait. To capture this correlation, we also implemented our method assuming that the joint expression and trait distribution is a full multivariate normal distribution. Unfortunately, the implementation of the multivariate normal and associated computations increases the computational time to an infeasible level. Thus, we do not use this approach in any of the results shown in this article. Despite this shortcoming, our simulations show that the method still performs well even in the presence of substantial extra-genotype correlation. There is likely some loss of power, but the false-positive rate is not increased (see RESULTS).

Using standard EM algorithm techniques, we derive estimators for the genotype-specific expression means and variances, genotype frequencies for the controlling locus, and genotype-specific clinical trait means and variances. Without the clinical trait data $\vec{Y}$, these would be the standard EM estimators for the normal mixture model. These estimators are modified to include the clinical trait data. See supporting information, File S1 for details.

Grouping genes: We wrote an algorithm that uses hierarchical clustering to assign transcripts into groups with similar expression patterns. We used the `hclust` function in R (<http://www.r-project.org>) with the expression data (clustering across genes) and a Euclidean distance and complete linkage to create a dendrogram relating the genes.

The algorithm starts by clustering the genes as described above. Then, the `cutree` function in R is used to divide the genes into two groups according to their position in the dendrogram. If the size of a group exceeds the target, then hierarchical clustering is applied again and the group is again divided into two smaller groups using `cutree`. This process is continued iteratively and groups are successively divided into smaller groups until no groups exceed the target size. Each of the current groups is checked in each iteration. If its size is less than or equal to the target size, then the group is added to a list of final groups. If its size is greater than the target, then it is divided into smaller groups. The code is included in File S1 with further documentation.

After the groups are selected, we apply our method to each one to test whether a controlling factor for that group is also a causative locus for the clinical trait. Note that we chose hierarchical clustering to produce the gene groups simply because it is a fast, simple method that produces results easy to separate into smaller groups. Our simulations show that it works adequately. However, other clustering methods may work better in some situations and our approach does not require any particular clustering method.

Determining significance: The statistical significance of the gene groups can be determined by using the asymptotic approximation for the likelihood ratio. There is a difference of two free parameters between the alternative and null hypotheses: three genotype-specific clinical trait means under the alternative and one genotype-independent clinical trait mean under the null. Therefore, we approximate the distribution of the likelihood ratio as an asymptotic chi-square distribution with 2 d.f. under the null hypothesis (note that the number of components in the mixture does not change between the null and the alternative hypothesis). Our simulations show that this approach works well. However, there is a potential problem associated with our assumption of independence between different transcripts after conditioning on genotype. This assumption will clearly not be met in some cases and

TABLE 1
Description of parameters in the simulations

Parameter	Explanation	Default values
L	No. of clinical trait QTL	See Figures 1–6
M	No. of transcripts with expression levels influenced by a given clinical trait QTL	See Figures 1–6
C	No. of clinical trait QTL influencing transcript	$1 \leq c \leq L$
P	Frequency of the allele causing an increase in the clinical trait	0.5
h_T	Dominance coefficient for the clinical trait	0.5
h_E	Dominance coefficient for the expression level	0.5
μ_{e0}, μ_{e2}	Mean expression levels when there are 0 and $2c$ clinical trait-increasing alleles	0, 1
μ_{t0}, μ_{t2}	Clinical trait mean for 0 and $2L$ trait-increasing alleles, respectively	0, 1
σ_E	Transcript and clinical trait standard deviation	0.333

there will be additional dependencies between transcripts. In this case the likelihood function is no longer correct and the asymptotic approximation for determining significance may not perform well. In particular, the probability of a set of transcripts having an apparent correlation with the clinical trait by chance may be underestimated. Our simulations (see [File S1](#)) show that the false-positive rate is not increased by this effect. However, because of the possibility that the false-positive rate could be affected with correlation structures different from what we tested, we feel that a permutation test is a more reliable means of assessing significance. Due to considerations of computational time, most of the simulation results that we show in this article use the asymptotic approximation to assess significance. However, we recommend using the permutation approach to assess significance with real data sets.

Local maxima: The EM algorithm is guaranteed only to find local maxima. We found that with a single selection of starting parameter values, the algorithm is quite likely to end up at a local maximum instead of the global one. We used two different approaches to avoid this. One is to choose multiple sets of random starting values for parameters, run the algorithm for each one, and take the parameter fits producing the highest likelihood as best fit. We found that 20 sets of starting values were usually adequate. The other approach we used was the split and merge EM (SMEM) algorithm (UEDA *et al.* 2000), which uses split and merge operations between different mixture components to find global maxima. This procedure had similar performance to 20 starting values in finding the global maximum, but was substantially faster computationally. Thus, most of our results use this approach.

We found the local maximum problem to be particularly serious in our application because we are using the EM algorithm to maximize likelihoods in a likelihood-ratio test. False positives can result if the EM algorithm finds the global maximum under the alternative hypothesis model but misses it under the null hypothesis one. Even using 20 starting values or the split-merge EM, the global maximum is still missed enough for this to be a serious concern. Giving the best-

fit parameters under the alternative hypothesis model as one of the sets of starting parameter values for the null hypothesis model greatly reduces this problem. We take the additional measure of conducting an exhaustive search of the parameter space for the null hypothesis model for gene groups that produce significant results, to ensure that the global maximum has not been missed. We do this by defining a grid of plausible values for each parameter and then trying all combinations of parameters over this grid as starting values for the EM.

Simulation study: We conducted simulations to evaluate our method for testing for association between expression levels and the clinical trait. For each individual, the genotypes ($g = 0, 1, 2$, clinical trait-increasing alleles) on L causative loci are randomly generated. The clinical trait is assumed to follow a normal distribution whose mean is the sum of effects contributed by each of the L loci. Trait means for genotypes 0, 1, and 2 on each locus are μ_{t0}/L , μ_{t1}/L , and μ_{t2}/L , respectively, where $\mu_{t1} = \mu_{t0} + h_T \times (\mu_{t2} - \mu_{t0})$ and h_T is the dominance coefficient. Each locus contributes equally to the clinical trait. Each transcript is assumed to follow a normal distribution with the mean being the sum of contributions from a subset of c of the L causative loci. Each subset of c causative loci is assumed to control the expression of M transcripts. Expression means for genotypes 0, 1, and 2 on each locus are μ_{e0}/c , μ_{e1}/c , and μ_{e2}/c , respectively, where $\mu_{e1} = \mu_{e0} + h_E(\mu_{e2} - \mu_{e0})$ and h_E is the dominance coefficient for expression.

All transcripts were simulated as independent after conditioning on the shared controlling genotype except where specifically noted otherwise. Likewise, the clinical trait was independent of the transcripts after conditioning on shared controlling genotype. The variance for both expression and the clinical trait is 0.1. This leads to heritability values for the transcripts of $\sim 38\%$ for $c = 1$, 12% for $c = 2$, and 6% for $c = 3$. The heritability for clinical trait QTL is $\sim 2\%$ for $L = 6$ and $\sim 1\%$ for $L = 9$ for the standard parameter values in this article.

Appropriate parameter values are largely unknown, so we vary these parameters over reasonable ranges. For

each set of parameter values, we simulated 1000 replicates to assess the power or false-positive rate in detecting the association between transcripts and the clinical trait. Table 1 gives definitions and default values of parameters used in the simulations.

QTL mapping: We use the JZmapqtl function in QTL Cartographer Version 1.17 (BASTEN *et al.* 1994) to map eQTL for gene groups. This function conducts composite-interval QTL mapping jointly for a set of traits on the basis of the method of JIANG and ZENG (1995). Their method assumes that traits follow a normal mixture model, similarly to our method. QTL Cartographer does not have functionality for QTL mapping for multiple gene groups. Thus, we wrote a Perl script that loops over each gene group, calls the JZmapqtl function to conduct joint QTL mapping, and collects information on significant QTL. This script is available in File S1. See the mouse data application for discussion on the choice of significance thresholds.

Data: We apply our method to the data of GHAZALPOUR *et al.* (2006). They conducted an F_2 cross between inbred mouse strains C3H/HeJ and C57BL/6J. Using 135 female offspring from this cross, they collected genotype information at 1065 SNPs, liver gene expression data, and data on 22 physiological traits, including body weight. From their microarray data, GHAZALPOUR *et al.* (2006) selected a set of the 3421 most varying and most connected genes to use for their gene network analysis. These data are publicly available and we obtained them from the authors' web site (<http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/MouseWeight/>).

RESULTS

Simulation study: We start with simulation results that show how the statistical power to detect that a transcript group shares a causative locus with the clinical trait depends on various parameter values and assumptions. Then, we show the application of our full genome-scan procedure to a mouse obesity data set.

Effect of gene group size on power: Figure 4 shows a plot of the power to detect that a set of n genes shares a controlling locus with the clinical trait. Specifically, it is the power to reject the null hypothesis that the controlling locus for the group of n transcripts is not a QTL for the clinical trait, as a function of sample size and other parameters. The group size n is shown next to each curve. In the simulations for Figure 4 it was assumed that genes were correctly grouped (*i.e.*, that all genes in the group share the same controlling locus). The n genes in the group were simulated such that they share $c = 3$ controlling loci with the clinical trait. Other parameters are as shown in Table 1. In general all parameter values in this article take the value shown in Table 1 unless otherwise specified. We see that the power increases quickly with group size. The sample size required for 80% power is >1000 for $n = 1$, ~ 900 for $n = 2$, 550 for

FIGURE 4.—Effect of transcript group size. The parameters are $L = 6$, $c = 3$, and otherwise as shown in Table 1. The y-axis is the power to detect that the transcript group is associated with the clinical trait. The x-axis is the total sample size. The number of transcripts in the group is shown for each curve. Results are based on 1000 replicates and power is assessed at a significance level of 0.05/1000.

$n = 4$, 450 for $n = 8$, and 350 for $n = 20$. As the number of genes in the group increases, the importance of a transcript's individual variation in expression is reduced and joint expression becomes a more reliable indicator of the underlying controlling genotype. Although most of the absolute improvement in power occurs in going from $n = 1$ to $\sim n = 4$, there is still substantial proportional decrease in required sample size going from $n = 12$ to $n = 20$ (and presumably beyond). This indicates that if (1) the group truly consists of transcripts sharing a controlling locus and (2) there is no correlation beyond that due to shared controlling genotype, then larger group size is always beneficial (at least out to some to unknown but large limit).

The simulations in Figure 4 assumed that the gene group always consists entirely of transcripts that are jointly controlled by the same clinical trait QTL. In reality, the gene groups are determined by hierarchical clustering (or another clustering method) with a preset group size. Thus, some transcript groups may contain transcripts that are not controlled by the QTL and this may adversely affect power. Results in File S1 show that these "extra" transcripts do not affect the results unless the number of transcripts not controlled by the clinical trait locus equals or exceeds the number controlled by that locus. Plots showing the impact of varying the dominance coefficients, trait means, and trait variances are also given in the Results section of File S1.

Effect of correlation between transcripts: Our method assumes that the only correlation between transcripts comes from shared genotype. That is, we assume that

FIGURE 5.—Plot of effect of correlation between transcripts beyond shared controlling genotype. In this plot, $L = 6$, $c = 3$, and group size is also 6. The additional correlations (beyond shared genotype) between transcripts are as shown in the plot. Results are based on 1000 replicates and power is assessed at a significance level of 0.05/1000.

the transcripts are independent after conditioning on genotype at the controlling loci. The simulation data from the above results follow this assumption. However, it is likely in reality that there will be additional correlation between transcripts. This correlation will decrease power. In the independent case, each transcript is giving an independent observation about the underlying genotype. When there is additional correlation between transcripts, then the amount of independent information is reduced. Figure 5 shows the effect of this correlation. When there was no correlation, expression values were generated with a common mean according to shared controlling genotype and no additional correlation. When, for example, the correlation is 0.2, then the expression values were generated from a multivariate normal distribution with a common mean and a pairwise correlation of 0.2 (that is, all of the off-diagonal terms in the covariance matrix are 0.2 times the product of the two expression standard deviations). The expression standard deviations all had the default value of 0.333.

The correlation due to shared genotype alone is 0.29 for the parameter values used in generating this plot. With additional correlations of 0.2, 0.4, and 0.6, the total correlations are 0.44, 0.58, and 0.72. Figure 5 shows that there is substantial effect of this additional correlation. For example, 80% power is reached at a sample size of ~ 400 with no extra correlation. The required sample size for 80% power increases to ~ 600 , 800, and 1000 when the extra correlations between transcripts are 0.2, 0.4, and 0.6, respectively. As this correlation gets very high, then power is actually reduced relative to the case using a single transcript because of the larger

number of parameters that must be fit. We see that power increases when there is a negative correlation between transcripts. When one transcript has a high value relative to its genotypic mean, then other transcripts in the group will tend to have low values and the deviations cancel each other out. It is not currently known how much of the correlation in expression between genes is typically due to shared genotype and how much is due to other types of shared control.

When there is additional correlation between transcripts, then our model is not correct and thus there is reason to be concerned that the asymptotic chi-square approximation may provide incorrect P -values. The probability of a set of transcripts having an apparent correlation with the clinical trait by chance may be underestimated because they are more highly correlated with each other than assumed by the model. Thus, we conducted simulations to test the false-positive rate when there is additional correlation. These results are available in [File S1](#) and show that the false-positive rate is not increased by this effect.

Comparison to standard QTL association mapping:

Figure 6 shows comparisons of our method to traditional QTL mapping for different combinations of the parameters c (the number of these QTL that jointly control expression of the transcript group), M (the number of transcripts that are jointly controlled by the c QTL), and L (the number of clinical trait QTL). The curves labeled “direct QTL mapping” give the power for directly testing for association between genotype at one of the L QTL and the clinical trait. This was calculated assuming that an ANOVA is being used to test for differences in the clinical trait between QTL genotypes. This gives an upper limit on power by standard QTL association mapping methods because we assume that the exact causative polymorphism is being tested. These curves were generated using the same parameter values as the other curves in the plot and also the same multiple-testing correction. We can use this to compare the power of standard QTL mapping to the power of our approach for testing whether a transcript group shares a QTL with the clinical trait. We see that when $c = 1$, the power achieved by our approach is slightly less than by direct QTL mapping, while the power of our method is much higher when $c > 1$. For example, a sample size of 1000 is required to achieve 80% power for direct QTL mapping with $L = 6$, while sample sizes of 700, 500, 400, and 350 are required for $c = 2, 3, 4$, and 5 with $M = 6$ for our approach. The power increases strongly with c . As c increases, then the transcript group and the clinical trait share more controlling loci and thus the expression patterns and the clinical trait become more concordant (this point is discussed extensively in SCHLIEKELMAN 2008). Standard QTL mapping is not able to benefit from this effect. Of course, our method requires an additional step of mapping the QTL for the transcript group, which is not modeled here. However, many

FIGURE 6.—Comparison of power with standard QTL mapping. The curve with diamonds shows the power to detect direct association between the clinical trait and the QTL genotype using an ANOVA. The other curves show the power for the expression-based approach. All parameter values were the same (were applicable) between the two approaches and as shown in Table 1 or on the plot. Results are based on 1000 replicates and power is assessed at a significance level of 0.05/1000.

studies have shown that the power to map such eQTL is often quite high (see reference citations in the Introduction). Furthermore, it is highly unlikely that the actual polymorphism would be tested (as opposed to some linked marker) unless the marker density is very high. In this case, the multiple-testing correction for QTL mapping would be much more stringent.

False-positive rate: Simulation results in File S1 show that the false-positive rate at a significance threshold of 0.05 is 5%, as expected. We have seen some cases of false positives caused by local maxima as discussed in METHODS. Thus, positive results should be carefully checked for this possibility (see METHODS).

In another example, we analyzed a data set that mixed data for simulated transcripts controlled by a clinical trait QTL with an expression data set published by CHANDRAN *et al.* (2007). The goal was to see whether we could detect the simulated clinical trait-related genes among the real genes. This allowed us to test the performance of our method in a complex data set with a realistic correlation structure in the background genes. The procedure performed well, detecting transcripts associated with each of the clinical trait QTL while producing only a single false-positive group out of 391. These results are in File S1.

Application of the full genome-scan procedure to a mouse obesity data set: Next, we show an application of the full genome-scan procedure to the data set of GHAZALPOUR *et al.* (2006). See METHODS for a description of the data. From their microarray data, GHAZALPOUR *et al.* (2006) selected a set of the 3421 most varying and most connected genes to use for their gene network analysis. We used the same set of

3421 genes. We applied hierarchical clustering and our gene grouping algorithm with a maximum group size of 6. This produced a total of 983 gene groups, with sizes ranging from 1 to 6.

We then applied our EM algorithm procedure to each gene group, using body weight as the clinical trait. We first used the asymptotic chi-square distribution with 2 d.f. with a Bonferroni correction to determine significance for the likelihood-ratio test. A total of 152 gene groups were found to have a controlling factor significantly associated with body weight. These gene groups contained a total of 576 genes. We also applied a permutation approach to assess significance, which gave similar results (File S1).

Next, we conducted QTL mapping using the R programming language (<http://www.r-project.org/>) with the R/qtl package (BROMAN *et al.* 2003). We first applied standard QTL mapping procedures to body weight, using permutations to assess significance. The four most significant SNP positions are shown in Table 2. There was one significant QTL at SNP p45500 and two suggestive QTL at SNPs p45558 and p44688.

Next, we mapped QTL for the 152 significant gene groups using the JZmapqtl function in QTL Cartographer. This function, based on the method of JIANG and ZENG (1995), treats a group of traits as a multivariate mixture model with components corresponding to genotypes at joint QTL and maps the QTL using composite-interval mapping. Its underlying model is similar to the one we assume for testing gene groups. The primary difference is that we infer genotype on the basis of expression data and Jiang and Zeng infer genotype using marker genotype data.

TABLE 2

The four most significant SNP positions from standard QTL mapping

SNP	Chromosome, position (Mb)	lod score	P-value
p45276	1, 142.9	3.03	0.243
p45558	5, 188.5	3.40	0.133
p44688	15, 76.7	3.58	0.098
p45500	19, 50.4	4.07	0.031

Following the suggestion of JIANG and ZENG (1995), statistical significance was assessed using a Bonferroni-corrected χ^2 -test with $2m + 1$ d.f., where m is the group size. We also used a permutation approach (see File S1) that gave similar results.

A total of 38 gene groups produced significant eQTL (Table 3) at a total of 45 different genomic locations. Some of these 45 SNPs are close to each other on the same chromosome and are thus apparently due to the same QTL. The first column in Table 3 shows groupings of significant SNPs into distinct loci. These groupings were based on visual inspection of the lod plots. Many of the peaks are broad and thus it is not clear in some cases whether significant SNPs are due to the same or different eQTL. The most ambiguous cases are marked in Table 3. The individual lod plots are given in File S1. By our judgment, there were 27 separate QTL.

We see that one significant SNP position, p44688, matches exactly with a suggestive SNP position from the standard QTL mapping of body weight. Furthermore, SNP p45328 is directly adjacent to significant SNP p45558 from the standard QTL mapping. The significant SNP p45500 from the standard QTL mapping is in the support interval of the significant SNP p45826 from our method. Finally, SNP p45276 that had the fourth lowest P -value (but not significant) in the standard QTL mapping is in the support interval of significant SNP p45924 from our method. Thus, our method produced significant QTL at all four of the most significant positions from the standard QTL mapping. Three of these four positions were significant by our method but not by the traditional gene mapping method. This gives a strong proof of principle for our approach.

There were 23 other significant QTL regions in addition to these four QTL. All but one of the significant QTL positions are mapped to the same genomic location as at least one gene in the group that produced the QTL. Thus, these genes become obvious candidate genes for the QTL. There were a total of 48 such *cis*-acting genes.

There are numerous instances of *cis*-eQTL for multiple genes mapping to the same location. These may be cases of genes being strongly linked to a body weight QTL, but whose own controlling genotype does not affect body weight. The expression levels of such genes could be significantly associated with body weight even though they do not affect it.

GHAZALPOUR *et al.* (2006) identified 12 transcription modules using gene coexpression networks constructed from the expression data. They then looked for correlations between these transcription modules and obesity-related clinical traits by calculating the average correlation between the traits and genes in the module. Finally, they mapped eQTL for all genes within the modules and looked for genomic regions that were enriched for eQTL for each module. They found nine such regions. Four of the nine peak markers were within the support interval of significant markers by our method (markers p45006, p46033, p45207, and p45826 from our method). Three others were not in the support interval of a significant marker, but within the same general genomic region as significant markers from our method (a distance of 30 Mb from marker p45328 and distances of 14 and 36 Mb from marker p44633). The remaining two significant regions from their method were on chromosome 10, where our method produced no significant markers. This strong overlap between the methods gives support for both.

We conducted literature review of the 48 candidate genes using the Entrez Gene database and PubMed to find information about their function. This information is summarized in Table 1 in File S1. Sixteen of the 48 genes have previously been associated with obesity or obesity-related traits. These 16 genes were divided among 11 of the putative QTL. That 16 of our candidate genes have been previously associated with obesity gives evidence that our method is producing biologically relevant results. That it also produced 16 more novel QTL shows that it potentially has great power to find causative loci for complex traits. Of course, it is not necessarily the case that a significant eQTL for a transcript associated with the clinical trait is causative for the clinical trait (see DISCUSSION).

Test of assumptions: Our method is based on several key assumptions. We use the four transcript groups that produced the eQTL located at or near the SNPs listed in Table 2 as case studies to examine the validity of these assumptions.

The most fundamental assumption is that polymorphisms that affect a clinical trait of interest at least sometimes also produce detectable changes in expression levels in some genes. We see strong evidence that is true for these data. The four polymorphisms showing the strongest relationship with body weight via traditional QTL mapping also produced strong *cis* expression changes. It is interesting to note that transcript group 18 that produced the eQTL near SNP p45500 (the only SNP that was significant by traditional QTL mapping) produced the highest lod score for expression linkage (23.6) and the 20th (of 983) lowest P -value for association with body mass. Thus, a strong relationship between genotype and the clinical trait is reflected in a strong relationship between genotype and expression and between expression and clinical trait.

TABLE 3
Significant transcripts with significant eQTL

Apparent locus	Gene group ^b	<i>cis</i> gene name	Chr, position (Mb)	eQTL SNP	Chr, position (Mb)
1	108	Igfbp2	1, 74.4	p46102	1, 111.5
1	17	Mogat1	1, 80.3	p44629	1, 100.4
1	326	Ngef	1, 89.4	p46102	1, 111.5
2	384	Adamts4	1, 175.1	p45924	1, 168.3
2	108	Prdc	1, 178.7	p45860	1, 172.8
3	107	Neb	2, 52.5	p45066	2, 55.3
4	518	Slc35c1	2, 93.0	p45546	3, 75.3
4b	799	Lhfp	3, 54.6	p45897	3, 50.0
4	422	Ppm1l	3, 71.1	p46033	3, 70.0
4	518	Fgb	3, 84.9	p44942	3, 79.4
5 ^a	97	Lrrc39	3, 119.9	p45832	3, 110.6
6	289	Cyp4b1, MMT00008970	4, 114.5	p45550	4, 105.9
7 ^a	208	2610002D18Rik	4, 133.4	p45246	4, 128.0
8	25	Spp1	5, 103.7	p45328	5, 112.5
8	143	Cxcl5, Niacr1	5, 91.2, 122.5	p45328	5, 112.5
9	440	Pparg ^c	6, 117.0	p46315	6, 53.1
9	138	Mug2	6, 123.8	p45846	6, 121.3
9	11	Osblp3	6, 50.8	p44480	6, 46.0
10 ^a	234	Pde6h ^c	6, 138.4	p45298	6, 138.6
11	194	—	—	p44616	7, 28.4
11	518	Ric3	7, 101.3	p44616	7, 28.4
12	483	Slco2b1	7, 92.1	p45571	7, 95.0
13	51	Defb1	8, 20.6	telomere	8, 8
13	704	Tusc3	8, 38.6	p45718	8, 28.6
13	154	Plat	8, 21.6	p44868	8, 21.2
14 ^a	308	Hice1	8, 71.5	p45109	8, 73.9
14 ^a	566	Ier2	8, 85.5	p45105	8, 60.8
15	559	Sparc	11, 55.0	p44783	11, 57.0
16	668	A830093I24Rik	12, 4.8	p44633	12, 2.2
17	400	2610204M08Rik	12, 108.2	p45657	12, 110.1
18	154	Serpina6b	13, 32.8	telomere	13, 0.01
19	380	Wdhd1	14, 39.0	p45207	14, 49.5
20	182	6330409N04Rik	14, 52.7	p44591	14, 56.0
21	349	BC025446	15, 77.1	p44688	15, 76.7
21	184	Lgals1	15, 80.3	p44794	15, 75.2
22	438	Gbe1	16, 71.2	p44923	16, 54.8
23 ^a	128	Acat2	17, 13.1	p45222	17, 5.9
24	483	Nfkbil1	17, 34.9	p44642	17, 42.3
24	325	Cyp4f15	17, 32.4	p45498	17, 31.9
25	383	Cd14	18, 36.8	p46168	18, 35.6
26 ^a	229	Tjp2	19, 24.1	p45312	19, 22.7
26 ^a	347	MMT00014892	19, 39.8	p45956	19, 28.1
27	108	Cyp2c37,Cyp2c54,Cyp2C50	39.7, 40.2, 40.2	p46074	19, 35.5
27	18	MMT00041314	19, 39.9	p46074	19, 35.5
27	18	Cyp2c40	19, 40.0	p45826	19, 37.2

^a There is ambiguity as to whether the QTL is distinct from the other QTL on the chromosome. b indicates that this gene may have a distinct QTL.

^b See Table Table S3 in File S1 for the genes in each group.

^c Significant as single gene trait only.

Another fundamental assumption of our method (although not of the general approach of using expression information in genome scans) is that genetic variation at QTL for the clinical trait drives variation in multiple transcripts (at least some of the time) and that those transcripts will therefore have highly correlated expression. The converse of this is that genes with

highly correlated expression share controlling loci at least some of the time.

Table 4 shows the genes in the 4 transcript groups that produced eQTL at or near the SNP positions given in Table 2. The groups contain 6, 2, 1, and 4 genes. Overall, the mean group size was 4.1 for the 48 significant groups (recall that the tree-cutting algorithm was set to produce

TABLE 4
Composition of transcript groups

Gene group	Gene name	Chromosome	Start position (Mb)	% lod below
18	<i>BC014805</i>	19	7.5	98.5
	<i>Cyp2g1</i>	7	20.8	92.3
	<i>CRAD-L</i>	3	1.5	100
	<i>Prhr</i>	15	10.4	98.1
	<i>MMT00041314^a</i>	19	39.9	100
	<i>Cyp2c40^a</i>	19	40.0	100
349	<i>1110055L24Rik</i>	2	60.5	90.1
	<i>BC025446^a</i>	15	77.1	100
25	<i>Spp1^a</i>	5	103.7	100
384	<i>2310057H16Rik</i>	18	23.5	70.2
	<i>Adams4^a</i>	1	175.1	100
	<i>Serpine1</i>	5	7.9	74.6
	<i>Mthfd2</i>	6	84.4	73.5

^a Genes that produced significant eQTL.

groups of size ≤ 6). Group 18 has 2 transcripts with significant eQTL near SNP p45500. The remaining 3 groups each have only a single transcript that produced a significant eQTL. However, although no other transcripts in these groups produced significant (over all marker and all transcripts) eQTL, there is evidence that most of these transcripts are influenced by the significant eQTL for the group. The last column in Table 4 shows the percentage of lod scores for the transcript that are below the maximum observed in the region of the significant eQTL for the group (the region is defined as all contiguous SNP positions with lod score > 5). We see that five of the six transcripts in group 18 had lod scores in the top 2% in the region of SNP p45500. The remaining transcript had a lod score value in that region in the top 8%. Thus, we see evidence that the body weight QTL is exerting *trans*-acting control on all of the 6 genes in the group. We see similar, although weaker, evidence for the transcripts in groups 349 and 284. Note that these percentages are not *P*-values and this analysis does not constitute a hypothesis test. We merely intend these results to be suggestive.

File S1 contains lod score plots for some of the 48 transcript groups for different chromosomes. Many transcript groups have many shared lod score peaks. We used a simulation approach to conduct a test of the hypothesis that these matches in lod peaks are due to chance. Each chromosome was divided into two equal parts and the standard deviation in the location of the lod peak across the transcript group in each region was determined. This was compared to the standard deviations in 1000 replicates in which lod peaks were placed randomly in each region. The null hypothesis was strongly rejected in many regions for most transcript groups (see File S1 for more detail).

We cannot tell anything from this analysis about the extent to which the transcripts in a group are indepen-

dently responding to genotypic variation. It could be, for example, that the transcripts in a group are tightly regulated by a single gene. In this case, the many shared lod score peaks could simply be due to variation (much of it random) in that controlling gene's lod score. On the other hand, similar-looking plots could also result if transcripts in a group share many eQTL of weak to moderate strength. We suspect that the truth is a combination of these two possibilities. It is beyond the scope of this article to explore these issues further.

DISCUSSION

Using both simulations and an application to data, we have demonstrated that integrating expression information into genome scans for complex traits can substantially increase power to detect complex trait loci relative to using traditional QTL mapping methods. There are three reasons for this. The first is that a transcript may share multiple causative loci with the clinical trait. In this case, the transcript is an indicator for a multilocus genotype and thus will have stronger genotypic concordance with the clinical trait than a single locus can. Essentially, such a transcript allows us to test association of a multilocus genotype with the clinical trait without having to test the many possible multilocus combinations.

The second reason is that we are, in effect, directly inferring association between the underlying controlling genotype and the clinical trait instead of using an indirect quantity such as allele sharing. It is well known that direct tests for association give good power to detect causative loci with common alleles such as we have modeled (RISCH and MERIKANGAS 1996). Our method has the added advantage that it collapses all alleles into expression-increasing and expression-decreasing. Thus, there are in effect only two allele types. Association tests have been

shown to suffer a major loss in power when there are multiple causative alleles (SLAGER *et al.* 2000). Our method behaves similarly to an association test that always has two alleles, regardless of the true number of alleles. See SCHLIEKELMAN (2008) for further discussion.

The third reason is that we are inferring the locus directly, not a marker. Thus, there is no recombination obscuring the relationship between genotype and trait. Of course, we have error resulting from variation in expression. In this sense, our method trades variation in recombination between marker and genotype for variation in expression. When there is just a single transcript, then this variation can be substantial. However, if there are multiple transcripts controlled by the locus, the effect of this variation can be reduced. Even with SNP-based studies, recombination can be substantial. For example, the Affymetrix GeneChip 500K SNP marker array produces an average correlation to the nearest tag SNP of 0.66–0.8, depending on the population (FRAZER *et al.* 2007). Results that will appear in a future publication show that we can infer individual genotypes with >90% accuracy with simulated data generated using similar parameter values to those in Table 1 and a gene group size of 6 (E. H. BREAZEL and P. SCHLIEKELMAN, unpublished data).

In the case of a cross with a dense marker map, the second and third benefits listed above will not apply. That is, there will be only two allele types and recombination between a QTL and the nearest marker will likely be minimal. There will still be a benefit in power if there are transcripts controlled by multiple causative loci. Furthermore, the use of dense markers requires a higher number of statistical tests and corresponding loss of power. There may be a benefit in using a less dense marker map and an expression-based approach like ours.

It is an open question as to the extent to which QTL identified by our method are QTL in the traditional sense. It is the case that all four significant and near-significant QTL from the traditional QTL mapping approach were significant by our approach. Thus, our method does capture “real” QTL. Furthermore, other significant QTL by our method have been previously identified as body weight QTL (File S1). However, it is not clear whether all of the loci that we identified would, for example, be identified as QTL in a sufficiently powerful traditional QTL study.

It should be emphasized that an eQTL for a transcript that is trait associated is not necessarily itself causative for the clinical trait. There are three ways that a transcript could be associated with a clinical trait. First, it could be causative for the trait. In this case, any eQTL for the transcript are necessarily causative for the clinical trait. Second, it could be reactive to a clinical trait QTL but not causative to the trait. In this case, it could have other eQTL that are not causative to the trait. The third possibility is that it is reactive to the

clinical trait. In this case, any QTL for the trait are also eQTL for the transcript. However, it could also have other eQTL that are not causative for the trait.

Assumptions of the method: The most fundamental assumption of our approach is that polymorphisms that contribute to variation in clinical traits also create consistent variation in the expression level of some set of genes (whether in *cis* or in *trans*). This assumption seems supported by our results using the data of GHAZALPOUR *et al.* (2006) as well other studies (see Introduction). To the extent that such consistent differences in expression do exist, then these differences can be used to infer genotype at the polymorphic locus or loci.

Random variation in expression level will obscure this genotypic information. A second fundamental assumption of our approach is that we can reduce the effect of this variation by combining information from multiple transcripts. There does seem to be little doubt that clinically important polymorphisms often do affect the expression of multiple transcripts. This is seen in our results and in various other studies (see references cited in the Introduction). There are multiple possible approaches to combining the information from these multiple transcripts. Our approach is to assume that there is some degree of independence between transcripts after conditioning on shared controlling genotype. If this is true, then each transcript gives an independent observation of the controlling genotype and we can combine this information to reduce the effect of individual transcript variation, as we have demonstrated here. It is not clear to what extent this assumption of conditional independence is true or how much is gained by combining information from multiple transcripts. Our simulations show (Figure 5) that there is a major benefit in power even when there is substantial correlation between transcripts beyond shared controlling genotype. However, we assumed a simple correlation structure and it is not clear what would happen with a more complicated structure containing feedback and more realistic patterns of connections between genes, etc. A major task for future work will be to begin to sort out the relationship between gene networks and genotypic variation and to develop statistical methods that take the true network structure into account.

In discussing assumptions, we must distinguish between assumptions of the simulations and assumptions of the method. The simulations are based on many assumptions about the relationship between genotype, expression, and the clinical trait, but these are not assumptions of the method. These include the assumptions of an additive trait model, dominance coefficients, allele frequencies, heritability of expression, and number of transcripts per controlling locus. Some plots are shown with key parameters varying, and other plots are available in File S1. The impact of these assumptions is discussed extensively in SCHLIEKELMAN (2008), and

most of those findings apply directly to the results in this study.

Comparison to other genetical genomics approaches: The unifying theme of recent genetical genomics studies is the idea that gene networks will be more strongly linked with clinical traits of interest than will individual genetic loci. Different studies have taken different approaches as to how to combine information from the genes in the network. Our approach is to posit the existence of a controlling locus or loci shared between the genes in a group and the clinical trait. We then use a maximum-likelihood approach to estimate the parameters of this locus. GHAZALPOUR *et al.* (2006), for example, identify gene groups associated with the clinical trait by using the average correlation coefficient between genes in the group and the trait. They then locate causative genomic regions by identifying eQTL hotspots for the significantly correlated gene groups. In another example, CHEN *et al.* (2008) identified gene subnetworks that contained large numbers of genes whose expression levels were identified as causatively associated and in *cis* with a known metabolic QTL region on mouse chromosome 1. The relative merits of these approaches will depend at least partially on what turns out to be true about the relationships between gene networks, causative genotypes, and clinical traits. For example, the average correlation coefficient is probably not a good summary of the relationship between a gene group and the clinical trait. Our approach may better use the available expression information and therefore have better power. On the other hand, our approach is based on specific assumptions about the relationship between gene networks and genotype and may be less robust to violations of these assumptions than the approach of GHAZALPOUR *et al.* (2006). The approach of GHAZALPOUR *et al.* (2006) identifies only causative regions that drive the expression of many genes in a trait-associated gene group. Similarly the approach of CHEN *et al.* (2008) and many other studies cited in the Introduction identify only networks with many genes mapping to a preidentified region of interest. While this will certainly produce interesting results, it may also miss important causative components. Further data will be required to determine which approach is best. Most likely, a combination of complementary approaches will be most effective.

This study was supported in part by the University of Georgia Research Computing Center, a partnership between the Office of the Vice President for Research and the Office of the Chief Information Officer of the University of Georgia.

LITERATURE CITED

- AYROLES, J. F., M. A. CARBONE, E. A. STONE, K. W. JORDAN, R. F. LYMAN *et al.*, 2009 Systems genetics of complex traits in *Drosophila melanogaster*. *Nat. Genet.* **41**: 299–307.
- BASTEN, C. J., B. S. WEIR and Z.-B. ZENG, 1994 Zmap-a QTL cartographer in *Computing Strategies and Software*, edited by C. SMITH, J. S. GAVORA, B. BENKEL, J. CHESNAIS, W. FAIRFULL *et al.* Organizing Committee, 5th World Congress on Genetics Applied to Livestock Production, University of Guelph, Guelph, Ontario, Canada, Vol. 22, pp. 65–66.
- BERGE, K. E., H. TIAN, G. A. GRAF, L. YU, N. V. GRISHIN *et al.*, 2000 Accumulation of dietary cholesterol in sitosterolemia caused by mutations in adjacent ABC transporters. *Science* **290**: 1771–1775.
- BREM, R. B., G. YVERT, R. CLINTON and L. KRUGLYAK, 2002 Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**: 752–755.
- BROMAN, K. W., H. WU, S. SEN and G. A. CHURCHILL, 2003 R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**: 889–890.
- BYSTRYKH, L., E. WEERSING, B. DONTJE, S. SUTTON, M. T. PLETCHER *et al.*, 2005 Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat. Genet.* **37**: 225–232.
- CHANDRAN, U. R., C. Q. MA, R. DHIR, M. BISCEGLIA, M. LYONS-WEILER *et al.*, 2007 Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process. *BMC Cancer* **7**: 64.
- CHEN, M., and C. KENDZIORSKI, 2007 A statistical framework for expression quantitative trait loci mapping. *Genetics* **177**: 761–771.
- CHEN, Y. Q., J. ZHU, P. Y. LUM, X. YANG, S. PINTO *et al.*, 2008 Variations in DNA elucidate molecular networks that cause disease. *Nature* **452**: 429–435.
- CHESLER, E. J., L. LU, S. SHOU, Y. QU, J. GU *et al.*, 2005 Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat. Genet.* **37**: 233–242.
- CHEUNG, V. G., L. K. CONLIN, T. M. WEBER, M. ARCARO, K. Y. JEN *et al.*, 2003 Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat. Genet.* **33**: 422–425.
- DEMPSTER, A. P., N. M. LAIRD and D. B. RUBIN, 1977 Maximum likelihood from incomplete data via Em algorithm. *J. R. Stat. Soc. Ser. B Methodol.* **39**: 1–38.
- EDWARDS, A. C., J. F. AYROLES, E. A. STONE, M. A. CARBONE, R. F. LYMAN *et al.*, 2009 A transcriptional network associated with natural variation in *Drosophila* aggressive behavior. *Genome Biol.* **10**: R76.
- EMILSSON, V., G. THORLEIFSSON, B. ZHANG, A. S. LEONARDSON, F. ZINK *et al.*, 2008 Genetics of gene expression and its effect on disease. *Nature* **452**: 423–428.
- FRAZER, K. A., D. G. BALLINGER, D. R. COX, D. A. HINDS, L. L. STUVE *et al.*, 2007 A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**: 851–861.
- GELFOND, J. A. L., J. G. IBRAHIM and F. ZOU, 2007 Proximity model for expression quantitative trait loci (eQTL) detection. *Biometrics* **63**: 1108–1116.
- GHAZALPOUR, A., S. DOSS, S. S. SHETH, L. A. INGRAM-DRAKE, E. E. SCHADT *et al.*, 2005 Genomic analysis of metabolic pathway gene expression in mice. *Genome Biol.* **6**: R59.
- GHAZALPOUR, A., S. DOSS, B. ZHANG, S. WANG, C. PLAISIER *et al.*, 2006 Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet.* **2**: 1182–1192.
- HSEIH, W. P., G. PASSADOR-GURGEL, E. A. STONE and G. GIBSON, 2007 Mixture modeling of transcript abundance classes in natural populations. *Genome Biol.* **8**: R98.
- HUBNER, N., C. A. WALLACE, H. ZIMDAHL, E. PETRETTO, H. SCHULZ *et al.*, 2005 Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat. Genet.* **37**: 243–253.
- JIA, Z. Y., and S. Z. XU, 2005 Clustering expressed genes on the basis of their association with a quantitative phenotype. *Genet. Res.* **86**: 193–207.
- JIANG, C. J., and Z. B. ZENG, 1995 Multiple-trait analysis of genetic mapping for quantitative trait loci. *Genetics* **140**: 1111–1127.
- KARP, C. L., A. GRUPE, E. SCHADT, S. L. EWART, M. KEANE-MOORE *et al.*, 2000 Identification of complement factor 5 as a susceptibility locus for experimental allergic asthma. *Nat. Immunol.* **1**: 221–226.
- KENDZIORSKI, C. M., M. CHEN, M. YUAN, H. LAN and A. D. ATTIE, 2006 Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics* **62**: 19–27.
- KIRST, M., A. A. MYBURG, J. P. G. DE LEON, M. E. KIRST, J. SCOTT *et al.*, 2004 Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA microarray

- data in an interspecific backcross of eucalyptus. *Plant Physiol.* **135**: 2368–2378.
- LAWN, R. M., D. P. WADE, M. R. GARVIN, X. WANG, K. SCHWARTZ *et al.*, 1999 The Tangier disease gene product ABC1 controls the cellular apolipoprotein-mediated lipid removal pathway. *J. Clin. Invest.* **104**: R25–R31.
- LI, H., L. LU, K. F. MANLY, E. J. CHESLER, L. BAO *et al.*, 2005 Inferring gene transcriptional modulatory relations: a genetical genomics approach. *Hum. Mol. Genet.* **14**: 1119–1125.
- MORLEY, M., C. M. MOLONY, T. M. WEBER, J. L. DEVLIN, K. G. EWENS *et al.*, 2004 Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**: 743–747.
- NETTLETON, D., and D. WANG, 2006 Selective transcriptional profiling for trait-based eQTL mapping. *Anim. Genet.* **37**: 13–17.
- OLEKSIK, M. F., G. A. CHURCHILL and D. L. CRAWFORD, 2002 Variation in gene expression within and among natural populations. *Nat. Genet.* **32**: 261–266.
- PEREZ-ENCISO, M., J. R. QUEVEDO and A. BAHAMONDE, 2007 Genetical genomics: use all data. *BMC Genomics* **8**: 69.
- RISCH, N., and K. MERIKANGAS, 1996 The future of genetic studies of complex human diseases. *Science* **273**: 1516–1517.
- SAMPSON, J. N., and S. G. SELF, 2008 Identifying trait clusters by linkage profiles: application in genetical genomics. *Bioinformatics* **24**: 958–964.
- SCHADT, E. E., 2006 A genome-wide association approach to mapping the genetic determinants of the transcriptome in human populations. *Eur. J. Hum. Genet.* **14**: 891–893.
- SCHADT, E. E., S. A. MONKS, T. A. DRAKE, A. J. LUSIS, N. CHE *et al.*, 2003 Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**: 297–302.
- SCHADT, E. E., J. LAMB, X. YANG, J. ZHU, S. EDWARDS *et al.*, 2005 An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.* **37**: 710–717.
- SCHADT, E. E., C. MOLONY, E. CHUDIN, K. HAO, X. YANG *et al.*, 2008 Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* **6**: 1020–1032.
- SCHLIEKELMAN, P., 2008 Statistical power of expression quantitative trait loci for mapping of complex trait loci in natural populations. *Genetics* **178**: 2201–2216.
- SILLANPAA, M. J., and N. NOYKOVA, 2008 Hierarchical modeling of clinical and expression quantitative trait loci. *Heredity* **101**: 271–284.
- SLAGER, S. L., J. HUANG and V. J. VIELAND, 2000 Effect of allelic heterogeneity on the power of the transmission disequilibrium test. *Genet. Epidemiol.* **18**: 143–156.
- UEDA, N., R. NAKANO, Z. GHAHRAMANI and G. E. HINTON, 2000 SMEM algorithm for mixture models. *Neural Comput.* **12**: 2109–2128.
- WAYNE, M. L., and L. M. MCINTYRE, 2002 Combining mapping and arraying: an approach to candidate gene identification. *Proc. Natl. Acad. Sci. USA* **99**: 14903–14906.
- YAN, H., W. YUAN, V. E. VELCULESCU, B. VOGELSTEIN and K. W. KINZLER, 2002 Allelic variation in human gene expression. *Science* **297**: 1143.

Communicating editor: H. ZHAO

GENETICS

Supporting Information

<http://www.genetics.org/cgi/content/full/genetics.110.123968/DC1>

A Genetical Genomics Approach to Genome Scans Increases Power for QTL Mapping

Guoying Sun and Paul Schliekelman

Copyright © 2011 by the Genetics Society of America
DOI: 10.1534/genetics.110.123968

FILE S1**SUPPORTING MATERIALS**

Explanation of EM algorithm – page 3

Tree cutting algorithm – page 6

Scripts for using QTL Cartographer for mapping gene groups – page 11

False positive rate – page 13

Application with mixed simulated and real data – page 16

Effect of transcripts in groups that are unrelated to the clinical trait – page 18

Effects of other simulation parameters – page 20

Determining significance thresholds with the Ghazalpour et al data – page 22

Permutation test for eQTL mapping with Ghazalpour et al data – page 24

Functional Information for the candidate genes in Ghazalpour data – page 28

Lod score plot for chromosomes with multiple eQTLs (Ghazalpour data) – page 38

Comparing lod score patterns within gene groups – page 47

EM Algorithm

We use the EM algorithm (Dempster et al. 1977) to obtain maximum likelihood estimates (MLE) of parameters since we do not actually observe the genotype. Matrix X and vector Y are treated as observed data. Genotype $\vec{G} = (g_1, g_2, \dots, g_N)^T$ is unobserved data, where $g_i = 0, 1, 2$, is the number of ELT-increasing alleles in individual i at the controlling locus. We assume that expression level distribution and the quantitative trait distribution are independent given the genotype at the controlling locus, and independence of expression levels within and between individuals given genotype and independence of the clinical trait status between individuals. Under H_1 , the complete data likelihood function is:

$$\begin{aligned} L(\Theta | \vec{X}, \vec{Y}, \vec{G}) &= f(\vec{X} | \vec{G}, \bar{\theta}) f(\vec{Y} | \vec{G}, \bar{\theta}) P(\vec{G} | \bar{\theta}) \\ &= \prod_{i=1}^N f(\bar{X}_i | g_i, \bar{\theta}) f(\bar{Y}_i | g_i, \bar{\theta}) P(g_i | \bar{\theta}) \end{aligned}$$

The first step of the EM algorithm, the E-step, is to obtain the expected value of the complete data log likelihood with respect to the unobserved data $\vec{G}$ given the observed data $\vec{X}, \vec{Y}$ and the current parameters estimates $\bar{\theta}_0$:

$$\begin{aligned} Q &= E_{\vec{G}}(\log L_1(\bar{\theta} | \vec{X}, \vec{Y}, \vec{G}) | \vec{X}, \vec{Y}, \bar{\theta}_0) \\ &= \sum_{i=1}^N \sum_{g_i} \log f(\bar{X}_i | g_i, \bar{\theta}) P(g_i | \vec{X}, \vec{Y}, \bar{\theta}_0) \\ &\quad + \sum_{i=1}^N \sum_{g_i} \log f(\bar{Y}_i | g_i, \bar{\theta}) P(g_i | \vec{X}, \vec{Y}, \bar{\theta}_0) \\ &\quad + \sum_{i=1}^N \sum_{g_i} \log P(g_i | \bar{\theta}) P(g_i | \vec{X}, \vec{Y}, \bar{\theta}_0) \\ &\doteq T_1 + T_2 + T_3 \end{aligned} \tag{2.1}$$

The second step of the EM algorithm iteration, the M-step, is to maximize the expected log likelihood with respect to the parameters $\bar{\theta}$. New parameter estimates will be produced from the M-step. We assume that the expression levels follow a normal distribution on some scale (e.g. log scale), and the clinical trait also follows a normal distribution. The first set of terms in (2.1) is maximized to obtain the parameters of the expression distribution.

$$\begin{aligned}
T_1 &= \sum_{i=1}^N \sum_{g_i} \log f(\bar{X}_i | g_i, \bar{\theta}) P(g_i | \bar{X}, \bar{Y}, \bar{\theta}_0) \\
&= \sum_{i=1}^N \sum_{g_i} \left[-\frac{1}{2} \log(|\Sigma_{x,g_i}|) - \frac{1}{2} (\bar{x}_i - \bar{\mu}_{x,g_i})^T \Sigma_{x,g_i}^{-1} (\bar{x}_i - \bar{\mu}_{x,g_i}) \right] P(g_i | \bar{X}, \bar{Y}, \bar{\theta}_0) \\
\hat{\mu}_{x,l} &= \frac{\sum_{i=1}^N \bar{X}_i P(g_i = l | \bar{X}_i, Y_i, \bar{\theta}_0)}{\sum_{i=1}^N P(g_i = l | \bar{X}_i, Y_i, \bar{\theta}_0)} \tag{2.2}
\end{aligned}$$

where $\hat{\mu}_{x,l} = (\mu_{1l}, \mu_{2l}, \dots, \mu_{nl})^T$, μ_{il} is mean for expression level i and genotype l .

$$\hat{\Sigma}_{x,l} = \frac{\sum_{i=1}^N P(g_i = l | \bar{X}_i, Y_i, \bar{\theta}_0) (\bar{X}_i - \hat{\mu}_{x,l})(\bar{X}_i - \hat{\mu}_{x,l})^T}{\sum_{i=1}^N P(g_i = l | \bar{X}_i, Y_i, \bar{\theta}_0)} \tag{2.3}$$

In practice, we constrain the off-diagonal terms to zero and use a product of independent normal pdfs for $f(\bar{x}_i | g_i, \hat{\theta})$. This is for computational reasons as discussed in the methods section in the main text.

The second term is maximized to find the parameters for the clinical trait given the genotype. It is also straightforward to show that:

$$\begin{aligned}
T_2 &= \sum_{i=1}^N \sum_{g_i} \log f(Y_i | g_i, \bar{\theta}) P(g_i | \bar{X}, \bar{Y}, \bar{\theta}_0) \\
&= \sum_{i=1}^N \sum_{g_i} \left(-\frac{1}{2} \log \sigma_{y,g_i}^2 - \frac{(y_i - \hat{\mu}_{y,g_i})^2}{2\sigma_{y,g_i}^2} \right) P(g_i | \bar{X}, \bar{Y}, \bar{\theta}_0) \\
\hat{\mu}_{y,l} &= \frac{\sum_{i=1}^N Y_i P(g_i = l | \bar{X}, \bar{Y}, \bar{\theta}_0)}{\sum_{i=1}^N P(g_i = l | \bar{X}, \bar{Y}, \bar{\theta}_0)} \tag{2.4}
\end{aligned}$$

$$\hat{\sigma}_{y,l}^2 = \frac{\sum_{i=1}^N (Y_i - \hat{\mu}_{y,l})^2 P(g_i = l | \bar{X}, \bar{Y}, \bar{\theta}_0)}{\sum_{i=1}^N P(g_i = l | \bar{X}, \bar{Y}, \bar{\theta}_0)} \tag{2.5}$$

The probability of genotype can be estimated by maximizing the third term as follows:

$$\hat{P}(g_i = l) = \frac{1}{N} \sum_{i=1}^N P(g_i = l | \bar{X}, \bar{Y}, \bar{\theta}_0) \tag{2.6}$$

The estimates in equation (2.2)-(2.6) are updated current parameters, to be used for the next EM iteration. In these equations, the distribution of the unobserved data genotype is needed. It can be written as:

$$\begin{aligned} P(g_i | \bar{X}, \bar{Y}, \bar{\theta}_0) &= \frac{f(\bar{X}, \bar{Y} | g_i, \bar{\theta}_0) P(g_i | \bar{\theta}_0) P(\bar{\theta}_0)}{f(\bar{X}, \bar{Y}, \bar{\theta}_0)} \\ &= \frac{f(\bar{X}_i | g_i, \bar{\theta}_0) f(Y_i | g_i, \bar{\theta}_0) P(g_i | \bar{\theta}_0)}{\sum_{g_i} f(\bar{X}_i | g_i, \bar{\theta}_0) f(Y_i | g_i, \bar{\theta}_0) P(g_i | \bar{\theta}_0)} \end{aligned}$$

Under H_0 , the estimates of the parameters of the expression distribution and the genotype probabilities have the same equations (2.2, 2.3 and 2.6) as under H_1 . However, the parameter estimates for clinical trait are different than under H_1 .

$$\hat{\mu}_y = \frac{\sum_{i=1}^N \sum_{g_i} Y_i P(g_i = l | \bar{X}, \bar{Y}, \bar{\theta}_0)}{\sum_{i=1}^N \sum_{g_i} P(g_i = l | \bar{X}, \bar{Y}, \bar{\theta}_0)} = \frac{\sum_{i=1}^N Y_i}{N} \quad (2.7)$$

$$\hat{\sigma}_y^2 = \frac{\sum_{i=1}^N \sum_{g_i} (Y_i - \hat{\mu}_y)^2 P(g_i = l | \bar{X}, \bar{Y}, \bar{\theta}_0)}{\sum_{i=1}^N \sum_{g_i} P(g_i = l | \bar{X}, \bar{Y}, \bar{\theta}_0)} = \frac{\sum_{i=1}^N (Y_i - \hat{\mu}_y)^2}{N} \quad (2.8)$$

The expression for the distribution of the unobserved data genotype under H_0 is also different from under H_1 :

$$P(g_i | \bar{X}, \bar{Y}, \bar{\theta}_0) = \frac{f(\bar{X}_i | g_i, \bar{\theta}_0) P(g_i | \bar{\theta}_0)}{\sum_{g_i} f(\bar{X}_i | g_i, \bar{\theta}_0) P(g_i | \bar{\theta}_0)}$$

Each iteration of the EM algorithm is guaranteed to increase the log-likelihood and the parameters converge to local maximum likelihood estimates (Redner and Walker 1984; Wu 1983; Xu and Jordan 1996).

Tree-cutting algorithm

The R code for the tree-cutting algorithm is shown below. Extensive explanation is given as comments. This code can be pasted into an R script file in order to run it. The basic steps of the program are outline below:

- 1) apply hierarchical clustering (across genes) to the expression data in `final_exp`.
- 2) Use function `cutree()` to divide the tree into two groups. Enter into the while loop. This loop continues as long as their are groups with group size > `ntarget`

Within the *while* loop, the *for* loop is over gene groups:

- 1) Take each group and check whether its size is $\leq ntarget$. If it is, then add the gene numbers to the vector `finalorder` and the group size to the vector `groups`.
 - 2) If the group size is greater than `ntarget`, then apply `hclust()` to the group and use `cutree` to cut the group into `clustgroup` (default 2) groups.
 - 3) The genes are ordered by their new grouping and the ordered gene numbers are concatenated to the vector `neworder`. Their group numbers are concatenated to the vector `B`. The group numbers are made unique for this while iteration by adding this previous total number of groups assigned (`tgrp`).
 - 4) At the completion of the *for* loop, the vector `neworder` has genes ordered by their grouping and the vector `B` has group numbers for each of these genes. This only contains genes that were in a group of size more than `ntarget` after the previous iteration.
 - 5) `neworder` and `B` are `cbind()` to make a matrix `newclust`. This matrix is the new gene matrix used in the next iteration of the while loop.
 - 6) The while loop continues as long as there is at least one group whose size is greater than `ntarget`.
 - 7) The final output is 1) `FinalEMdata_h1`, which is the expression data (with missing data removed) reordered by gene groups; 2) `Order_h1`, which is a vector of orderings mapped back to the original gene indexing 3) the vector `groups`, which gives the sizes for each group. We must combine the ordering and the group sizes to reconstruct the groups.
-

R Code:

#This program reads in expression data and clinical trait data. It deals with missing data and then partitions the genes into groups of size ntarget or less, according to shared expression patterns.

#rm(list=ls()) #clear all variable names

#User-set parameters.

clustgroup<-2

ntarget=6 #maximum number of genes in each group

#Read in the data:

exp_femalef2<-read.csv("LiverExpression.csv",stringsAsFactors=F) # read expression data into R

exp_femalef2<-exp_femalef2[,-c(1:8,144:150)]

exp_femalef2<-as.matrix(exp_femalef2)

r<-nrow(exp_femalef2) #get the number of row

c<-ncol(exp_femalef2) #get the number of column

genodat<-read.csv("F2mouseBXHxC3HGenotypes.csv",header=F,stringsAsFactors=F)

weight<-genodat[1,-(1:3)]

weight<-as.numeric(weight)

#scan(file="phenotype_weight_femalef2.txt") # read weight

#Deal with missing expression data-----

#replace the missing data with mean of the row(gene) if there are less than 10% are missing,

#otherwise, delete the gene

num_NA<-apply(is.na(exp_femalef2),1,sum) # obtain the number of missing values for each gene

*exp_female<-exp_femalef2[num_NA<c*0.1,] #remove rows with more than 10% missing values*

*genes<-(1:r)[num_NA<c*0.1] #vector of non-discarded genes*

rowmeans<-apply(exp_female,1,mean,na.rm=T) #calculate row means with missing values removed

#replace each na with the row mean:

for(i in 1:nrow(exp_female))

{ exp_female[i,is.na(exp_female[i,])]=rowmeans[i]

}

#

#Deal with missing weight data:-----

```
f_weight<-weight[!is.na(weight)] # delete missing values
final_exp<-exp_female[!is.na(weight)] #delete columns corresponding to missing values
```

```
rr<-nrow(final_exp)
```

```
#
```

*#Next, we divide the genes into groups of size $\leq n_{target}$ by their gene expression similarity.
#The algorithm works as follows:*

```
# 1) apply hierarchical clustering (across genes) to the expression data in final_exp
# 2) Use function cutree() to divide the tree into two groups.
#Enter into the while loop. This loop continues as long as their are groups with group size>n_target
# Within the while loop, the for loop is over gene groups:
#     1) Take each group and check whether its size is  $\leq n_{target}$ . If it is, then add the gene numbers
#         to the vector finalorder and the group size to the vector groups.
#     2) If the group size is greater than n_target, then apply hclust() to the group and use cutree
#         to cut the group into clustgroup (default 2) groups.
#     3) The genes are ordered by their new grouping and the ordered gene numbers are concatenated
#         to the vector neworder. Their group numbers are concatenated to the vector B. The group
#         numbers are made unique for this while iteration by adding this previous total number of
#         groups assigned (tgrp).
#     4) At the completion of the for loop, the vector neworder has genes ordered by their grouping and
#         the vector B has group numbers for each of these genes. This only contains genes that were in
#         a group of size more than n_target after the previous iteration.
#     5) neworder and B are cbind() to make a matrix newclust. This matrix is the new gene matrix used in the
#         # next iteration of the while loop.
#     6) The while loop continues as long as there is at least one group whose size is greater than n_target.
#     7) The final output is 1) FinalEMdata_h1, which is the expression data (with missing data removed)
reordered
#         by gene groups; 2)Order_h1, which is a vector of orderings mapped back to the original gene
indexing
#         3) the vector groups, which gives the sizes for each group. We must combine the ordering and the
group
#         sizes to reconstruct the groups.
```

#Note that this code is inefficient as R code and thus is probably slower than it needs to be.

```
# clustering the genes
clustdata<- final_exp
order11<-(1:rr) #index of genes
orgorder<-cbind(order11,genes) # relates the current (reduced) gene index to the original (full) gene index
```

```
xclust<-hclust(dist(clustdata)) # hierarchical clustering of the genes
```

#we want to zoom in on the clusters at the bottom of the tree:

```
xcut<-cutree(xclust,k=clustgroup) # cut the tree into k groups
cluster<-cbind(order11,xcut) # 1st column is the order, 2nd cloumn is group number
```

```
# sort cluster by group
```

```
newclust<-cluster[order(cluster[,2]),] # sort by group, get the corresponding order
```

```
ngrp<-clustgroup
```

```
order1<-newclust[,1] # the order after first cut
```

```
groups<-NULL # will be a vector containing the number of genes in each group
```

```
finalorder<-NULL #will be a vector listing genes in order by their groupings
```

```
# replicate until every group has equal to or less than maximum number of genes
```

```
while(length(newclust[,2])>=1)
```

```
{ num=0 #will store total number of genes counted
```

```
  tgrp=0 #will store total number of groups assigned so far
```

```
  B<-NULL # B is the next column of newclust
```

```
  neworder<-NULL
```

```
  for( i in 1:ngrp) #loop over current groups
```

```
  {
```

```
    sum1=sum(newclust[,2]==i) #calculate how many genes in the group i
```

```
    num=num+sum1 #total number of genes counted so far
```

```
    # if the group has gene number <= target number, then don't cut it again
```

```
    # store the genes and corresponding order:
```

```
    if(sum1<=ntarget){
```

```
      groups<-c(groups,sum1)
```

```
      # the number of genes in each group
```

```
      finalorder<-c(finalorder,order1[(num-sum1+1):num]) # the order corresponding to each group
```

```
    } else {
```

```
    # If the group i has more than a target number of genes, cut it again:
```

```
    #row<-order1[(num-sum1+1):num]
```

```
    row<-newclust[newclust[,2]==i,1] # rows of genes in the group to be cut
```

```
    redata<-clustdata[row,] # get the gene expression values for the genes in the group
```

```
    xclust1<-hclust(dist(redata)) # hierarchical clustering of group
```

```
    xcut1<-cutree(xclust1,k=clustgroup) # cut into k groups
```

```
    cluster1<-cbind(row,xcut1) # gives group number for each gene
```

```
    # Order the genes by their assigned group. The ordered gene numbers are concatenated
```

```
    # to the vector neworder. The group number is concatenated to the vector B. This groups number
```

```
    # is made unique for this iteration of the while loop by adding tgrp (the total number of previously
```

```
    # assigned groups) to it.
```

```
    neworder<-c(neworder,cluster1[order(cluster1[,2]),1]) # store the new order after sorting
```

```
    B<-c(B,sort(cluster1[,2])+tgrp) # accumulate the group number,
```

```
    # e.g. in one loop, there are 2 groups that have more than
```

```
    # target number genes, then next loop, each group will be cut
```

```
    # into 2 groups, so total will 4 groups
```

```
    tgrp<-tgrp+clustgroup #tgrp is the total number of groups assigned in current while
```

```
    # iteration
```

```
  }
```

```
} #end i loop (over groups)
```

```
newclust<-cbind(neworder,B) #this gives group numbers for all genes that haven't yet received final
```

```
  # group assignment.
```

```
ngrp<-tgrp
```

```
#this is total number of groups assigned for this iteration of while.
```

```
order1<-neworder
```

```
#contains the ordering of all remaining genes. used in next while
```

```
iteration.
```

```
}
```

*#At the end of the while loop, then the vector finalorder contains an ordering of all genes by groups and the
vector groups contains the group sizes for each group. We need these two together in order to reconstruct groups.*

```
cumgroup_h1<-cumsum(groups) #cumulative sum of group sizes
FinalEMdata_h1<-clustdata[finalorder,] #this takes the expression data (with missing data removed) and
                                         #reorders it according to the gene groups
Order_h1<-orgorder[finalorder,2] #this maps the final ordering of genes back to the original indexing
```

#

#Output-----

```
write(nrow(FinalEMdata_h1),file="fmice_parameters_n6h1.txt",ncol=1,append=T)
write(length(groups),file="fmice_parameters_n6h1.txt",append=T)
write(groups,file="fmice_parameters_n6h1.txt",ncolumns=length(groups),append=T)
write(cumgroup_h1,file="fmice_parameters_n6h1.txt",ncolumns=length(cumgroup_h1),append=T)

write(t(FinalEMdata_h1),file="fmice_finalEMdata_n6h1.txt",ncolumns=ncol(FinalEMdata_h1),append=F)
write(cumgroup_h1,file="fmice_cumgroup_n6h1.txt",ncolumns=length(cumgroup_h1),append=F)
write(groups,file="fmice_group_n6h1.txt",ncolumns=length(groups),append=F)
write(Order_h1,file="fmice_order_n6h1.txt",ncolumns=length(Order_h1),append=F)
write(f_weight,file="weight_old.txt",ncolumns=length(f_weight),append=F)
```

Scripts for using QTL Cartographer for mapping gene groups.

QTL Cartographer does not have functionality for mapping multiple gene groups from a microarray data set. Rather, separate runs must be conducted manually for each gene group. This is not feasible in our case because of the large number of gene groups. Thus, we wrote a Perl script to read in gene groups, create the necessary input files, call QTL Cartographer to do the mapping and collate the results. This script is shown below. A separate R script that is called by the Perl script is also given.

Perl script:

```
#this program does the following: 1) reads gene groups into $ggfile; 2) loops
over gene groups;
#3) for each gene group creates a crossfile for QTL cartographer (containing
header information, marker
#genotype information, and expression information); 4) calls QTL Cartographer
functions Rcross and JZmapQTL to
# do the QTL mapping for each gene group; 5) Calls R program findmax.R to
find the marker with the max lod score
# for the current gene group and prints this to a file called maxmark.txt;
Steps 3)-5) are repeated for each gene
# group. The final output is the file written to from $outfile. This contains
the info on the marker with the
# max lod score for each gene group.

use strict;
use warnings;

open(my $outfile, '>', "output/maxmark.txt"); #will contain marker with max
lod for each gene group
open(my $alllods, '>', "output/all_lods.txt");
open(my $ggfile, '<', "../inputfiles/gene_groups.txt") or die "Can't open
input file!: $!"; # gene groups
open(my $expfilein, '<', "../inputfiles/liver_expression_tabel.txt") or die
"Can't open input file!: $!"; #exp data
open(my $genofile, '<', "genotypes.txt") or die "Can't open input file!: $!";
#genotype data

my @expfile=<$expfilein>; #read in expression data
my @geno=<$genofile>; #read in genotype data

my $gcount=0;
while(<$ggfile>){ #this loop consecutively reads each line of $ggfile into
the default $_

    $gcount=$gcount+1;
    system("rm qtlcart.*"); #delete previous qtl cartographer files
    open(my $crossout, '>', "crossfile.txt");

    #lines below write required info to cross file for QTL Cartographer
    print $crossout "# 23232 -filetype cross.inp\n";
    print $crossout "-Cross RF2\n";
```



```

print $crossout "-otraits 0\n";
print $crossout "-SampleSize 135\n";
print $crossout "-missingtrait NA\n";
print $crossout "-case yes\n";
print $crossout "-TranslationTable\n";
print $crossout "AA 2 A\n";
print $crossout "Aa 1 H\n";
print $crossout "aa 0 B\n";
print $crossout "A- 12 12\n";
print $crossout "a- 10 10\n";
print $crossout "-- -1 -\n";

#read in gene group
my @gg=split; #splits input string into seperate parts
my $nt= $#gg+1; #number of genes in group

print $crossout "-traits $nt\n";

print $crossout @geno; #print genotype information to cross file

print $crossout "-start traits\n";

foreach my $ggv (@gg) {printf $crossout $expfile[$ggv];} #writes the
expression information for the gene group                                     #to $crossout

# foreach my $ggv (@gg) {printf "$expfile[$ggv]\n";}

print $crossout "-stop traits\n";
print $crossout "-quit\n";
print $crossout "-end\n";

#Call QTL Cartographer function Rcross to reformat data. Writes to
ghazalpour_mapout.map
system("Rcross -A -V -m ghazalpour_mapout.map -i crossfile.txt -n 135 -c
RF2 ");
#Call QTL Cartographer function JZmapqtl to do joint qtl mapping.
system("JZmapqtl -A -V -m ghazalpour_mapout.map -t 100 -I 30");
#Call R program findmax.R to take the QTL Cartographer output qtlcart.zo
to find the max lod position for
# the gene group and write it to the file maxmark.txt.
system("R CMD BATCH findmax.R");
system("cat qtlcart.z0 >> output/all_lods.txt") #saves all qtl
cartographer output for joint gene group.
#Comment out if you don't want
this saved.

}

```

Program findmax.R

```

dat1<-read.table("qtlcart.z0", skip=27,fill=T, as.is=c(T,T,T,T,T,T)) #fill=T
because of last line ("-e")

```

```
write.table(dat1[which.max(dat1[,5]),],file="output/maxmark.txt",append=T,row
.names=F,col.names=F)
```

False positive rate

We tested the false positive rate for our approach in several ways. First, we directly tested the EM algorithm with simulated data for gene groups that were unrelated to the clinical trait. The tables below show the uncorrected number of false positives for 1000 replicates for several combinations of parameter values for the simulated data. The false positive rate does not appear to deviate from the expected 5%.

False Positive Rate

Table 1. L=6 M=6 c=1

Sample Size	200	300	400	500
False Positives	49/1000	46/1000	42/1000	59/1000

Table 2. L=6 M=3 c=1

Sample Size	200	300	400	500
False Positives	60/1000	45/1000	67/1000	42/1000

Table 3. L=6 M=10 c=1

Sample Size	200	300	400	500
False Positives	43/1000	40/1000	57/1000	46/1000

We also tested the false positive rate in an application of our method (including screening, clustering, and hypothesis testing via the EM algorithm) to a data set constructed by combining a real microarray data set with a simulated clinical trait and associated transcripts. This allowed us to test the false positive rate with more realistic correlation structure. We found the false positive rate to be as expected. This is described in detail in the section ***Example Application with Mixed Simulated and Real Microarray Data.***

Effect of correlation on false positive rate. The above results used data that was simulated in a way that satisfied the assumptions of our methods. However, there is a potential problem associated with our assumption of independence between different transcripts after conditioning on genotype. This assumption will clearly not be met in many cases and there will be additional

dependencies between transcripts. In this case the likelihood function is no longer correct and the asymptotic approximation for determining significance may not perform well. In particular, the probability of a set of transcripts having an apparent correlation with the clinical trait by chance may be underestimated. In order to examine this effect, we tested the false positive rate with data simulated so that there was a correlation between genes even after conditioning on genotype. Typical results are shown below. The data was simulated with a multivariate normal with the off-diagonal terms (i,j) in the covariance matrix equal to $(\text{correlation}) * \text{SD}(\text{gene } i) * \text{SD}(\text{gene } j)$.

There does not appear to be any deviation from the expected false positive rate of 5%.

Table 4. L=6 M=6 c=1 Sample size=500

Correlation	-0.2	0	0.2	0.4	0.6	0.8
False Positives	40/1000	39/1000	50/1000	55/1000	51/1000	51/1000

Table 5. L=6 M=10 c=1 Sample size=500

Correlation	0	0.2	0.4	0.6	0.8
False Positives	62/1000	44/1000	55/1000	52/1000	40/1000

Example Application with Mixed Simulated and Real Microarray Data

In order to further test our method, we simulated a data set for a clinical trait and a set of jointly controlled ELTs similarly to above and then mixed it with the microarray data set of Chandran et al (2007). The goal is to see whether we can detect the simulated clinical trait related genes among the real genes. This allows us to test the performance of our in a complex data set with a realistic correlation structure in the background genes. The data set has expression data for 12,553 genes for each of 87 prostate cancer patients and 77 normal individuals. It is available on line at <http://www.ncbi.nlm.nih.gov>. We only used the 77 normal individuals and didn't include the prostate cancer status in our analysis. Instead, we simulated a clinical trait with $L=6$ QTLs and 200 ELTs that were also controlled by these six QTLs. Other parameters are as shown in Table 1, resulting in a heritability of about 2% for the QTLs. We simulated 200 ELTs: 50 controlled by cQTLs 1-3, 50 controlled by loci 4-6, 50 controlled by loci 1-2, and 50 controlled by loci 3-4. We generated 600 sample individuals in a manner similar to that described previously. For each simulated individual we randomly sampled an individual with replacement from the real data set and added the expression values from their 12553 genes to those from the 200 simulated genes (resulting in a total of 12753 expression measurements for each individual). We added a small amount of random noise to each of the real gene measurements to prevent their being many identical expression values in the data. This procedure preserves the correlation structure from the microarray data set. We then applied the screening and clustering procedure to the data set. We randomly selected 100 individuals from the data set and used this subset to calculate the correlation coefficient between each of the 12753 genes and the clinical trait. The 950 genes with the highest correlation with the clinical trait were retained and the rest discarded from further analysis. 92 out of 200 of the simulated genes made it through this screen.

The 950 selected genes were clustered into small groups with similar genes using hierarchical clustering our tree-cutting algorithm as previously described. Finally, we applied our EM algorithm method to all the groups to see whether we can detect the simulated genes.

The clustering procedure was conducted for maximum group sizes of 4 and 12 as shown. With a maximum group size of four, there were 414 groups total with 391 being comprised of real genes from the data set and 23 being comprised of simulated cQTL-controlled genes. With a maximum group size of 12, there were 166 groups with 156 being comprised of real genes from the data set and 10 being comprised of simulated cQTL-controlled genes. No groups in either case mixed cQTL-controlled and non-cQTL-controlled genes. The EM algorithm procedure was

then applied to each group. Table 2 shows the results. A total of 13 out of 23 cQTL-controlled groups were found to be significant with group size of four. Twelve out of the thirteen ELT groups (containing 38 ELTs) controlled by three cQTLs were found to be associated with the trait, but only one out of ten of the ELT groups (containing 3 ELTs) controlled by two cQTLs were found significant. When the maximum group size was twelve, five out of five $c=3$ groups (40 ELTs) were detected and two out of five $c=2$ groups (19 ELTs) were detected. One out of 391 non-cQTL groups were falsely detected as associated with the clinical trait with a group size of four. There were no false positives with a group size of twelve.

TABLE S1**Results from mixed data**

group size	Controlling Loci	Selected by Screen		Significant	
		Groups	ELTs	groups	ELTs
4	cQTL 1-3	11	32	10	30
	cQTL 4-6	2	8	2	8
	cQTL 1-2	9	27	1	3
	cQTL 3-4	1	3	0	0
	Non-cQTL	391	908	1	1
12	cQTL 1-3	4	32	4	32
	cQTL 4-6	1	8	1	8
	cQTL 1-2	4	27	2	19
	cQTL 3-4	1	4	0	0
	Non-cQTL	156	908	0	0

This shows the number of ELT groups and the number of ELTs that were selected in the screening step and were found as significantly associated with the clinical trait. Results are divided according to the controlling loci for the ELT group as shown.

Effect of transcripts in groups that are unrelated to the clinical trait: The results in Figure 4 show that power always increases with group size. However, the simulations in Figure 4 assumed that the gene group always consists entirely of transcripts that are jointly controlled by the same clinical trait QTL. In reality, the gene groups are determined by hierarchical clustering (or other clustering method) with a preset group size. Thus, some transcript groups may contain transcripts that are not controlled by the QTL. As the groups size increases, so does the probability of including such transcripts. Figure 5 shows the effect of such “extra” transcripts on power. The notation d12n9 means, for example, that the group contained 12 transcripts controlled by the same clinical trait QTL and 9 transcripts controlled by other loci unassociated with the clinical trait. There is no discernable effect on the power unless the number of transcripts in the group controlled by such loci equals or exceeds the number of transcript controlled by clinical trait QTLs. There is, at most, a very small decrease in power between the d12n0, d12n6, and d12n9 curves (the stochastic variations in the curves obscure the effects somewhat). Likewise, there is no discernable effect change in power between d6n0 and d6n3. However, there is a substantial decrease in power going to d6n6 and d6n9. For example, 80% power is achieved at a sample size of about 500 for d6n3, while a sample size of about 600 is required for 80% power for d6n9. If number of transcripts controlled by a clinical trait QTL exceeds the number of transcripts controlled by other QTLs, then apparently the parameter trajectory of the EM algorithm is dominated by the clinical trait QTL. This ceases to be true when the number of other transcripts equals or exceeds the number controlled by the clinical trait QTL. It should be noted that all transcripts had the same parameter values in our simulations. When parameter values vary between transcripts, then higher heritability ones are likely to impact the parameter trajectory more than lower heritability ones.

FIGURE S1.—Effect of non-clinical-trait-related transcripts on power. In this plot, $L=6$ and $c=3$. The curves correspond to different transcript group compositions as shown. For example for d12n6, the group has 12 transcripts controlled by a clinical trait QTL and 6 transcripts controlled by a locus unrelated to the clinical trait. Results are based on 1000 replicates and power is assessed at a significance level of 0.05/20000.

Effects of other parameters

Impact of dominance coefficients on power

To investigate the impact of dominance coefficients on the tests of association between ELTs and the clinical trait, we simulated data using different dominance coefficients for the clinical trait and ELTs, while other parameters are the same for each case as shown in table 2.1. When $h_E = 0$, expression level trait is recessive and there is no mean difference between individuals with genotype 0 and 1 at the causative locus. The overall ELT mean difference is bigger, which results in good power. Figure S2 shows that the power increases as the dominance coefficient for the clinical traits (h_T) decreases when the dominance coefficients for ELTs (h_E) are 0; and when $h_T = 0$, the power for $h_E = 0$ is much better than that for $h_E = 0.5$. These cases are the same as expectation. The power is the best when both ELTs and the clinical trait are recessive (i.e., $h_E = 0$ and $h_T = 0$). But when $h_E = 0.5$, the power is the same for $h_T = 0, 0.5$ and 1.

FIGURE S2.—Impact of dominance coefficients for expression and the clinical trait on power. In this plot, $L=6$, $M=6$ and $c=2$. Results are based on 1000 replicates. Power is assessed at level 0.05/1000.

Impact of trait mean μ_{t0} on power

Figure S3 displays the effect of μ_{t0} on power. As expected, the power decreases as μ_{t0} increases. For example, the power is 80% at sample size 800 for $\mu_{t0} = 0.1$, but the power is 65% at the sample size 800 for $\mu_{t0} = 0.2$ and it requires sample size 1000 to obtain the power 80%. For $\mu_{t0} = 0.5$, the power is very poor: the power is about 16% even at sample size 1000. For $\mu_{t0} = 0$, the trait means are $0/L$, $0.5/L$ and $1/L$ for genotype 0, 1 and 2 on each disease locus for $h_T = 0.5$. While for $\mu_{t0} = 0.2$, the trait means are $0.2/L$, $0.6/L$ and $1/L$ for genotype 0, 1 and 2 on each disease locus. Therefore, the trait mean difference decreases as μ_{t0} increases, then the power is smaller for bigger μ_{t0} .

FIGURE S3.—Impacts of clinical trait mean. In this plot, $L=6$, $M=6$ and $c=2$. Results are based on 1000 replicates and power is assessed at level of 0.05/1000.

Determining significance thresholds with the Ghazalpour et al data.

As discussed in the methods section, there is some reason to doubt whether asymptotic chi-squared approximation for the likelihood ratio is valid. Thus, we also used a permutation test as an alternative approach to assessing significance. The body weight values were randomly permuted between individuals and the EM algorithm was applied to produce a log-likelihood ratio value for each gene group. This was repeated 1000 times and the maximum log-likelihood ratio over the 983 gene groups was retained from each permutation. The 95% quantile of these values was taken as the significance threshold. Initially, this procedure produced a very high significance threshold of 22.5 and no gene groups were significant. Examination of the permutation results revealed that a single gene group (consisting of genes MMT00053504, MMT00041987, and MMT00080321) produced the highest likelihood ratio value in 144 out of 1000 permutations. A simple simulation procedure shows that the probability is less than 1 in 10,000 that any one of 983 gene groups should produce the top likelihood ratio more than 10 times in 1000 permutations by chance. Further investigation found that this group apparently has a strong tendency for the EM algorithm to converge to a local maximum instead of the global one under the null hypothesis model. This results in a highly inflated likelihood ratio value because the EM algorithm does not find global maximum the alternate hypothesis model. Elimination of this gene group reduced the significance threshold to 15.4. With this threshold, a total of 113 gene groups containing 432 genes were found to be significant. When eQTLs for these genes were mapped, we found 20 significant markers. Thus, eight genes that produced significant eQTLs were not found to be associated with body weight when we used the permutation approach.

The significance threshold using the asymptotic chi-squared approximation with the Bonferroni correction was 9.9. Thus, using the permutation test led to a very substantial increase in the significance threshold. This is the opposite of what we normally observe. That is, permutation tests usually decrease the significance threshold because they account for correlations between tests. Detailed examination of our results suggests that this increase results from occasional failures of the EM algorithm to find the global maximum under the null hypothesis model. Convergence to local maxima was also occasionally observed for the alternate hypothesis model, as evidenced by negative values for the log-likelihood ratio. Because of this effect, we feel that the permutation test leads to an overly stringent significance threshold. On the other hand, we cannot be certain that the asymptotic chi-squared approximation produces a

sufficiently stringent significance threshold because of the possibility of errors in the likelihood function due to correlations between genes after conditioning on genotype (see *Determining Significance* section). However, the lowest log-likelihood ratio value among the 28 gene groups that produced significant eQTLs was 12.7. Thus, we feel that all of these are likely to be truly significant.

Given the possibility of the EM algorithm not finding the null hypothesis global maximum likelihood value, there is the possibility of false positives among the 28 gene groups producing significant eQTLs. To avoid this, we conducted exhaustive searches of the parameter space for those gene groups under the null hypothesis model to ensure that the global maximum was located. This did not produce a loss of significance for any group.

Permutation test for eQTL mapping

In addition to doing joint eQTL mapping for gene groups, we also did single trait QTL mapping using each of the 576 genes in significant groups as phenotypes. We used the R/qtl package (BROMAN *et al.* 2003) with permutations to determine an overall significance threshold. In each permutation, the body weight values were permuted between sample individuals and then QTL mapping was conducted over all 576 genes. The maximum lod score over all genes and all markers was stored in each permutation and the upper 5% quantile of these values was taken as the significance threshold. Thus, this gives a significance threshold that accounts for multiple testing in both number of genes and in number of markers. We used 3000 permutations and a 5% significance threshold of 6.6. A total of 28 genes produced lod scores that exceeded this threshold. No gene had more than one significant eQTL. Figure S4 shows a plot of lod scores for these 28 genes over the genome and Table S2 details the 28 genes.

We see that the results are similar to those with the joint QTL mapping, although there are fewer significant gene groups and eQTLs. This is expected, since mapping gene groups jointly adds information over single-trait QTL mapping and thus power will often be higher. Two genes, *Pparg* and *Pde6h* were significant as single traits but were not part of gene groups that produced significant eQTLs in the joint analysis. These are included as candidate genes in Table 3 in the main text.

It is notable that even though the joint mapping used a Bonferroni-corrected significance threshold and the single trait mapping used permutations (which is usually substantially more liberal), the joint mapping produced more significant markers. This shows the effectiveness of joint mapping for this type of data.

FIGURE S4.—Plot of lod scores for the 28 significant transcripts. The x-axis is the ordered index for the 1065 markers (and not actual position). The horizontal line shows the cutoff for significance as determined from permutations over all markers and transcripts.

TABLE S2
Significant transcripts with significant eQTLs

Apparent Locus	Gene Name	Chr, Position (Mb)	eQTL SNP	Chr, Position(Mb)	Lod score
1	Igfbp2	1,74.4	p46102	1,111.5	6.6
1	Mogat1	1,80.3	p44629	1,100.4	9.6
1	Ngef	1,89.4	p46102	1,111.5	6.8
2	Adamts4	1,175.1	p45924	1,168.3	9.9
2	Prdc	1,178.7	p45860	1,172.8	6.6
3	Neb	2,52.5	p45066	2,55.3	10.0
4	Slc35c1	2,93.0	p45546	3,75.3	7.3
4b	Lhfp	3,54.6	p45897	3,50.0	10.0
4	Ppm1l	3,71.1	p46033	3,70.0	11.2
5	Spp1	5,103.7	p45328	5,112.5	8.0
6	Pparg	6,117.0	p46315	6,53.1	7.8
6	Mug2	6,123.8	p45846	6,121.3	8.2
7*	Pde6h	6,138.4	p45298	6,138.6	7.4
8	Slco2b1	7,92.1	p45571	7,95.0	19.7
9*	Tusc3	8,38.6	p45718	8,28.6	6.8
10	Hice1	8,71.5	p45109	8,73.9	8.0
11	A830093I24Rik	12,4.8	p44633	12,2.2	6.9
12	2610204M08Rik	12,108.2	p45657	12,110.1	11.1
13	Wdhd1	14,39.0	p45207	14,49.5	15.7
13	6330409N04Rik	14,52.7	p44591	14,56.0	17.0
14	BC025446	15,77.1	p44688	15,76.7	12.7
15	Gbe1	16,71.2	p44923	16,54.8	7.3
16*	Acat2	17,13.1	p45222	17,5.9	9.9
17	Nfkbil1	17,34.9	p44642	17,42.3	13.9
18	Cd14	18,36.8	p46168	18,35.6	7.7
19*	Tjp2	19,24.1	p45312	19,22.7	11.3
20	MMT00041314	19,39.9	p46074	19,35.5	14.8
20	Cyp2c40	19,40.0	p45826	19,37.2	23.6

*indicates that there is ambiguity as whether the QTL is distinct from the other QTL on the chromosome.

b indicates that this gene may have a distinct QTL.

TABLE S3**Genes in Significant Gene Groups**

Group number	Gene Indexes
108	1978 1999 2122 2503 2798 2824
17	1695 2084 2505
326	2720 2848
384	1233 2463 2687 3125
107	3405
518	1061 1367 1373 1432 1616 2669
799	213 926 1235
422	599 614 624 818 843 1922
97	822, 1692, 1896, 2130
289	823, 2452, 2580, 3411
208	801, 1024, 1244, 3264
25	2415
143	1044, 1983, 2853
440	1861 2067 2168 2707 2718 2934
138	2380 2990 3327
11	1139, 1450, 2777
234	3048 3198
194	692, 737, 1027, 1868
51	640, 1308, 1320, 1453, 2180, 3053
704	637 782 935 1313 1481 2026
154	1292, 1390, 1519, 2000, 2311, 2348,
308	1813 2038
566	2253, 3088, 2298
559	533, 669
668	1072 2252 2743
400	1700 2595 2750 2951
380	994 1270 1325 1350 2764
182	877 1524 1957 2123 2559 2693
349	2868 2942
184	2298, 2511, 2667
438	1955 1965 2085 2400 3224 3391
128	725 2020 2244
483	740 962 3065 3320
325	1750, 2148, 2324, 2383, 2425
383	655 1106 1838 1913 2265
229	1420 1564 2094 2291 3307 3353
347	2033, 2057, 2485, 2543, 3081
18	1818 1879 2438 2579 2723 3122

Gene indexes correspond to the rows in the expression data file provided by Ghazalpour et al. (2006) (<http://www.genetics.ucla.edu/labs/horvath/>)

Functional Information for the Candidate Genes

The table below summarizes functional information for the 28 genes selected by our method as candidate genes for body weight.

Apparent Locus	Gene Symbol/Name	Functional Information	Previously associated with obesity or obesity-related traits ?
1	Igfbp2 insulin-like growth factor binding protein 2	“has a critical role, either directly or indirectly, in modulating spleen and liver size” ² Expression level associated with weight gain ¹⁴	Yes ¹⁴
1	MGAT1 monoacylglycerol acyltransferase	“In the intestine, MGAT plays a major role in the absorption of dietary fat” ¹	Yes ²⁹
1	Ngef neuronal guanine nucleotide exchange factor	“These findings reveal that ephexin1 plays a role in axon guidance and is regulated by a switch mechanism that is specifically tailored to control Eph-mediated growth cone collapse” ⁵	No
2	Adamts4 a disintegrin-like and metallopeptidase (reprolysin type) with thrombospondin type 1 motif, 4	“female ADAMTS-4/5-double-knockout mice had a lower mean terminal body weight at the 12-week time point. Eight weeks after surgical induction of joint instability, osteoarthritis was significantly less severe in ADAMTS-4/5-double-knockout mice compared with WT mice.” ³	Yes ³
2	Prdc gremlin 2 homolog, cysteine knot superfamily	“PRDC is a potent BMP antagonist with a wide tissue expression pattern, and ovarian PRDC expressed in granulosa cells could be involved in follicular development...” ⁴	No
3	Neb Nebulin	“Nebulin is a giant modular sarcomeric protein that has been proposed to play critical roles in myofibrillogenesis, thin filament length regulation, and muscle contraction.” ⁷	Candidate gene ¹⁴
4	Slc35c1 solute carrier family 35, member C1	“Slc35c1-deficient mice presented with severe growth retardation, elevated postnatal mortality rate, dilatation of lung alveoles, and hypocellular lymph nodes.” ⁶	Yes ⁶
4b	Lhfp lipoma HMGIC fusion partner	provisional – uncharacterized in mouse ¹⁸ human: “we have isolated a novel human gene, LHFP (lipoma HMGIC fusion partner), that acts as a translocation partner of HMGIC in a lipoma” ⁸	No
4	Ppm1l protein phosphatase 1	“Three genes in this network, lipoprotein lipase (Lpl), lactamase beta (Lactb) and protein phosphatase 1-like (Ppm1l), are validated as previously unknown obesity genes, strengthening the association between this network and metabolic disease traits.” ⁹	Yes ⁹
4	Fgb fibrinogen beta chain	“The protein encoded by this gene is the beta component of fibrinogen, a blood-borne glycoprotein comprised of three pairs of	Yes ³⁶

		nonidentical polypeptide chains. Following vascular injury, fibrinogen is cleaved by thrombin to form fibrin which is the most abundant component of blood clots. In addition, various cleavage products of fibrinogen and fibrin regulate cell adhesion and spreading, display vasoconstrictor and chemotactic activities, and are mitogens for several cell types. Mutations in this gene lead to several disorders, including afibrinogenemia, dysfibrinogenemia, hypodysfibrinogenemia and thrombotic tendency.” ²⁰	
5	Lrrc39	functionally uncharacterized	No
	Cyp4b1	“This gene encodes a member of the cytochrome P450 superfamily of enzymes. The cytochrome P450 proteins are monooxygenases which catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids.” ²⁰	No
6			
6	MMT00008970	uncharacterized	No
7	2610002D18Rik	uncharacterized	No
	Cxcl5 chemokine (C-X-C motif) ligand 5	“CXCL5 or epithelial neutrophil activating peptide (ENA-78) is a cytokine belonging to the family of chemokines that is mainly implicated in the chemotaxis of inflammatory cells through the generation of local concentration gradients” ³¹ “CXCL5 is dramatically increased in serum of human obese compared to lean subject” ³⁰	Yes ³⁰
8			
	Niacr1 niacin receptor 1	“The orphan G-protein-coupled receptor GPR109B [Niacr1] is the result of a recent gene duplication of the nicotinic acid and ketone body receptor GPR109A being found in humans but not in rodents. Like GPR109A, GPR109B is predominantly expressed in adipocytes and is supposed to mediate antilipolytic effects.”	No
8			
	Spp1 secreted phosphoprotein 1 upregulated in obesity	“These data indicate a role for OPN in tissue remodeling in vivo, and suggest physiological functions during matrix reorganization after injury” ¹⁰	Yes ¹¹
8		“Osteopontin (OPN) is a multifunctional protein involved in various inflammatory processes, cell migration, and tissue remodeling....In high-fat diet-induced and genetically obese mice, OPN expression was drastically up-regulated in AT (40 and 80-fold, respectively) but remained largely unaltered in liver (<2-fold).” ¹¹	
	Pparg peroxisome proliferator activated receptor gamma	“The peroxisome proliferator-activated receptor gamma (PPARgamma) mediates the activity of the insulin-sensitizing thiazolidinediones and plays an important role in adipocyte differentiation and fat accretion.” ¹²	Yes ¹²
9			

9	Mug2 murinoglobulin 2	“Murinoglobulins are the single chain members of the alpha 2-macroglobulin family of proteinase inhibitors in the mouse.” ¹⁶	No
9	Osblp3 oxysterol binding protein-like 3	“This gene encodes a member of the oxysterol-binding protein (OSBP) family, a group of intracellular lipid receptors. Most members contain an N-terminal pleckstrin homology domain and a highly conserved C-terminal OSBP-like sterol-binding domain. Several transcript variants encoding different isoforms have been identified.” ²⁰	No
10	Pde6h phosphodiesterase 6H, cGMP-specific, cone, gamma	“In rod photoreceptors of wild-type mice, background light produces an acceleration of the decay of responses to brief flashes, accompanied by a decrease in the rate-limiting time constant for response decay. In rods in which phosphodiesterase gamma (PDEgamma) lacks one of its sites of phosphorylation (T35A rods), both the waveform of response decay and the rate-limiting time constant are nearly unaffected by backgrounds.” ¹³	No
11	Ric3 resistance to inhibitors of cholinesterase 3 homolog	Part of KEGG murine metabolism pathway “RIC3 is a protein associated with nicotinic acetylcholine receptors (nAChRs), neurotransmitter-gated ion channels expressed at the neuromuscular junction and within the central and peripheral nervous systems, that can enhance functional expression of multiple nAChR subtypes” ²⁰	No
12	Slco2b1 solute carrier organic anion transporter family, member 2b1	“the organic anion transporting polypeptide 2B1 (OATPB/SLCO2B1) is expressed in many organs, including liver, placenta, mammary gland, brain, and intestine. However, little is known about its function in those tissues....” ¹⁷	No
13	Defb1 defensin beta 1	“Defensins form a family of microbicidal and cytotoxic peptides made by neutrophils. Members of the defensin family are highly similar in protein sequence. This gene encodes defensin, beta 1, an antimicrobial peptide implicated in the resistance of epithelial surfaces to microbial colonization. This gene maps in close proximity to defensin family member, defensin, alpha 1 and has been implicated in the pathogenesis of cystic fibrosis.” ²⁰	No
13	Tusc3 tumor suppressor candidate 3	provisional – uncharacterized in mouse ¹⁸ Human: “This gene is a candidate tumor suppressor gene. It is located within a homozygously deleted region of a metastatic prostate cancer. The gene is expressed in most nonlymphoid human tissues including prostate, lung, liver, and colon. Expression was also detected in many epithelial tumor cell lines. Two transcript variants encoding	No

		distinct isoforms have been identified for this gene.” ²⁰	
13	Plat plasminogen activator, tissue	“This gene encodes tissue-type plasminogen activator, a secreted serine protease which converts the proenzyme plasminogen to plasmin, a fibrinolytic enzyme. Tissue-type plasminogen activator is synthesized as a single chain which is cleaved by plasmin to a two chain disulfide linked protein. This enzyme plays a role in cell migration and tissue remodeling. Increased enzymatic activity causes hyperfibrinolysis, which manifests as excessive bleeding; decreased activity leads to hypofibrinolysis which can result in thrombosis or embolism. Alternative splicing of this gene results in multiple transcript variants encoding different isoforms.” ²⁰	No
13	Hice1 HEC1/NDC80 interacting, centrosome associated 1	provisional – uncharacterized in mouse ¹⁸	No
14	Ier2 immediate early response 2	“shows a rapid and long-lasting induction after cerebral ischemia; induces cell death in primary neurons and displays several hallmarks of pro-apoptotic activity upon overexpression” ^{32,33}	No
15	SPARC secreted protein, acidic, cysteine-rich (osteonectin)	“Secreted protein acidic and rich in cysteine/osteonectin/BM40, or SPARC, is a matrix-associated protein that elicits changes in cell shape, inhibits cell-cycle progression, and influences the synthesis of extracellular matrix” ²⁰	Yes ³⁴
16	A830093I24Rik	predicted – uncharacterized ¹⁸	No
17	2610204M08Rik	“Is a WASP homology 2 motif-containing formin that severs actin filaments and accelerates both polymerization and depolymerization” ¹⁹	No
18	Serpinb6b serine (or cysteine) peptidase inhibitor, clade B, member 6b	“The serpin superfamily of serine protease inhibitors is implicated in the regulation of numerous physiological processes. In mice, Spi3/Serpinb6 has a broad tissue distribution.” ³⁵	No
		“Serpins are thought to regulate the activity of proteases in a wide variety of processes such as coagulation, fibrinolysis, complement activation, fertilization, inflammation, malignancy, tissue remodeling, and apoptosis....” ³⁵	
19		provisional – uncharacterized in mouse ¹⁸	No
		Human: “The protein encoded by this gene contains multiple N-terminal WD40 domains and a C-terminal high mobility group (HMG)	

	Wdhd1 WD repeat and HMG-box DNA binding protein 1	box. WD40 domains are found in a variety of eukaryotic proteins and may function as adaptor/regulatory modules in signal transduction, pre-mRNA processing and cytoskeleton assembly. HMG boxes are found in many eukaryotic proteins involved in chromatin assembly, transcription and replication. Alternative splicing results in two transcript variants encoding different isoforms. [provided by RefSeq] ²⁰	
20	6330409N04Rik	predicted – uncharacterized ¹⁸ human ortholog is candidate for B-cell chronic lymphocytic leukemia ²¹	No
21	BC025446	Model only - uncharacterized ²¹	No
21	Lgals1 lectin, galactose binding, soluble 1	“The galectins are a family of beta-galactoside-binding proteins implicated in modulating cell-cell and cell-matrix interactions. This gene product may act as an autocrine negative growth factor that regulates cell proliferation.” ²⁰	No
22	Gbe1 glucan (1,4-alpha-), branching enzyme 1	Provisional – uncharacterized in mouse Human: “The protein encoded by this gene is a glycogen branching enzyme that catalyzes the transfer of alpha-1,4-linked glucosyl units from the outer end of a glycogen chain to an alpha-1,6 position on the same or a neighboring glycogen chain. Branching of the chains is essential to increase the solubility of the glycogen molecule and, consequently, in reducing the osmotic pressure within cells. Highest level of this enzyme are found in liver and muscle. Mutations in this gene are associated with glycogen storage disease IV (also known as Andersen's disease). [provided by RefSeq] ²⁰ ”	Yes ²²
23	Acat2 acetyl-Coenzyme A acetyltransferase 2	“ACAT2 is a transmembrane protein that is expressed in the rough endoplasmic reticulum of enterocytes and esterifies free cholesterol that has been either synthesized <i>de novo</i> or internalized from the intestinal lumen” ²³	Yes ^{24,25}
24	Nfkbil1 nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor-like 1	“This gene encodes a divergent member of the I-kappa-B family of proteins. Its function has not been determined. The gene lies within the major histocompatibility complex (MHC) class I region on chromosome 6. Multiple transcript variants encoding different isoforms have been found for this gene.” ²⁰	No
24	Cyp4f15 cytochrome P450, family 4, subfamily f, polypeptide 15	No functional information available. Part of KEGG metabolism pathway	No
25	Cd14 CD14 antigen	“CD14 is a surface protein preferentially expressed on monocytes/macrophages. It	Yes ²⁶

		binds lipopolysaccharide binding protein and recently has been shown to bind apoptotic cells. Alternative splicing results in multiple transcript variants encoding the same isoform. [provided by RefSeq] ²⁰	
		“mice harboring mutations in TLR4 or in CD14, a co-receptor for TLR4, have an "ideal" body plan consisting of increased bone mineral content, density, and size as well as decreased body fat” ²⁶	
26	Tjp2 tight junction protein 2	“Tight junction proteins (TJPs) belong to a family of membrane-associated guanylate kinase (MAGUK) homologs that are involved in the organization of epithelial and endothelial intercellular junctions. TJPs bind to the cytoplasmic C termini of junctional transmembrane proteins and link them to the actin cytoskeleton.[supplied by OMIM] ²⁰	No
26	MMT00014892	no information available	No
27	Cyp2c37 cytochrome P450, family 2 subfamily c, polypeptide 37	“This gene encodes a member of the cytochrome P450 superfamily of enzymes. The cytochrome P450 proteins are monooxygenases which catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids.” ²⁰	Yes ²⁷
27	Cyp2c54 cytochrome P450, family 2, subfamily c, polypeptide 54	See previous	No (KEGG metabolic pathway)
27	Cyp2c50 cytochrome P450, family 2, subfamily c, polypeptide 54	See previous	No (KEGG metabolic pathway)
27	MMT00041314	no information available	No
27	Cyp2c40 cytochrome P450, family 2, subfamily c, polypeptide 40	“We conclude that CYP2C40 is one of the major cytochrome P450 proteins in the mouse intestinal tract. In the light of vasoactive and anti-neutrophilic effects of 16-HETE, we hypothesize that CYP2C40 may play an important role in endogenous biological functions in intestine.” ²⁸	Yes ²⁷
Part of KEGG metabolism pathways			

¹Yen CL, Stone SJ, Cases S, Zhou P, Farese RV Jr. Identification of a gene encoding MGAT1, a monoacylglycerol acyltransferase. Proc Natl Acad Sci U S A. 2002 Jun 25;99(13):8512-7

²Wood TL, Rogler LE, Czick ME, Schuller AG, Pintar JE. Selective alterations in organ sizes in mice with a targeted disruption of the insulin-like growth factor binding protein-2 gene. Mol Endocrinol. 2000 Sep;14(9):1472-82.

- ³Majumdar MK, Askew R, Schelling S, Stedman N, Blanchet T, Hopkins B, Morris EA, Glasson SS. Double-knockout of ADAMTS-4 and ADAMTS-5 in mice results in physiologically normal animals and prevents the progression of osteoarthritis. *Arthritis Rheum*. 2007 Nov;56(11):3670-4
- ⁴Avsian-Kretchmer O, Wang LS, Hsueh AJ. Protein related to DAN and cerberus is a bone morphogenetic protein antagonist that participates in ovarian paracrine regulation. *J Biol Chem*. 2004 May 28;279(22):23134-41
- ⁵Sahin M, Greer PL, Lin MZ, Poucher H, Eberhart J, Schmidt S, Wright TM, Shamah SM, O'connell S, Cowan CW, Hu L, Goldberg JL, Debant A, Corfas G, Krull CE, Greenberg ME. Eph-dependent tyrosine phosphorylation of ephexin1 modulates growth cone collapse. *Neuron*. 2005 Apr 21;46(2):161-3.
- ⁶Hellbusch CC, Sperandio M, Frommhold D, Yakubenia S, Wild MK, Popovici D, Vestweber D, Gröne HJ, von Figura K, Lübke T, Körner C. Golgi GDP-fucose transporter-deficient mice mimic congenital disorder of glycosylation IIc/leukocyte adhesion deficiency II. *J Biol Chem*. 2007 Apr 6;282(14):10762-72.
- ⁷Bang ML, Li X, Littlefield R, Bremner S, Thor A, Knowlton KU, Lieber RL, Chen J. Nebulin-deficient mice exhibit shorter thin filament lengths and reduced contractile function in skeletal muscle. *J Cell Biol*. 2006 Jun 19;173(6):905-16.
- ⁸Petit MM, Schoenmakers EF, Huysmans C, Geurts JM, Mandahl N, Van de Ven WJ. LHFP, a novel translocation partner gene of HMGIC in a lipoma, is a member of a new family of LHFP-like genes. *Cancer Genet Cytogenet*. 2002 Feb;133(1):90-3.
- ⁹Chen Y, Zhu J, Lum PY, Yang X, Pinto S, MacNeil DJ, Zhang C, Lamb J, Edwards S, Sieberts SK, Leonardson A, Castellini LW, Wang S, Champy MF, Zhang B, Emilsson V, Doss S, Ghazalpour A, Horvath S, Drake TA, Lusis AJ, Schadt EE. Variations in DNA elucidate molecular networks that cause disease. *Nature*. 2008 Mar 27;452(7186):429-35.
- ¹⁰Liaw L, Birk DE, Ballas CB, Whitsitt JS, Davidson JM, Hogan BL. Altered wound healing in mice lacking a functional osteopontin gene (spp1). *J Clin Invest*. 1998 Apr 1;101(7):1468-78.
- ¹¹Kiefer FW, Zeyda M, Todoric J, Huber J, Geyeregger R, Weichhart T, Aszmann O, Ludvik B, Silberhumer GR, Prager G, Stulnig TM. Osteopontin expression in human and murine obesity: extensive local up-regulation in adipose tissue but minimal systemic alterations. *Endocrinology*. 2008 Mar;149(3):1350-7.
- ¹²Imai T, Takakuwa R, Marchand S, Dentz E, Bornert JM, Messaddeq N, Wendling O, Mark M, Desvergne B, Wahli W, Chambon P, Metzger D. Peroxisome proliferator-activated receptor gamma is required in mature white and brown adipocytes for their survival in the mouse. *Proc Natl Acad Sci U S A*. 2004 Mar 30;101(13):4543-7.

- ¹³Woodruff ML, Janisch KM, Peshenko IV, Dizhoor AM, Tsang SH, Fain GL. Modulation of phosphodiesterase6 turnoff during background illumination in mouse rod photoreceptors. *J Neurosci*. 2008 Feb 27;28(9):2064-74.
- ¹⁴Hoeflich A, Wu M, Mohan S, Föll J, Wanke R, Froehlich T, Arnold GJ, Lahm H, Kolb HJ, Wolf E. Overexpression of insulin-like growth factor-binding protein-2 in transgenic mice reduces postnatal body weight gain. *Endocrinology*. 1999 Dec;140(12):5488-96.
- ¹⁵Ghazalpour A, Doss S, Zhang B, Wang S, Plaisier C, Castellanos R, Brozell A, Schadt EE, Drake TA, Lusis AJ, Horvath S. Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet*. 2006 Aug 18;2(8):e130.
- ¹⁶Overbergh L, Hilliker C, Lorent K, Van Leuven F, Van den Berghe H. Identification of four genes coding for isoforms of murinoglobulin, the monomeric mouse alpha 2-macroglobulin: characterization of the exons coding for the bait region. *Genomics*. 1994 Aug;22(3):530-9.
- ¹⁷Grube M, Köck K, Karner S, Reuther S, Ritter CA, Jedlitschky G, Kroemer HK. Modification of OATP2B1-mediated transport by steroid hormones. *Mol Pharmacol*. 2006 Nov;70(5):1735-41.
- ¹⁸Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, Ishii Y, Arakawa T, Hara A, Fukunishi Y, Konno H, Adachi J, Fukuda S, Aizawa K, Izawa M, Nishi K, Kiyosawa H, Kondo S, Yamanaka I, Saito T, Okazaki Y, Gojobori T, Bono H, Kasukawa T, Saito R, Kadota K, Matsuda H, Ashburner M, Batalov S, Casavant T, Fleischmann W, Gaasterland T, Gissi C, King B, Kochiwa H, Kuehl P, Lewis S, Matsuo Y, Nikaido I, Pesole G, Quackenbush J, Schriml LM, Staubli F, Suzuki R, Tomita M, Wagner L, Washio T, Sakai K, Okido T, Furuno M, Aono H, Baldarelli R, Barsh G, Blake J, Boffelli D, Bojunga N, Carninci P, de Bonaldo MF, Brownstein MJ, Bult C, Fletcher C, Fujita M, Gariboldi M, Gustincich S, Hill D, Hofmann M, Hume DA, Kamiya M, Lee NH, Lyons P, Marchionni L, Mashima J, Mazzarelli J, Mombaerts P, Nordone P, Ring B, Ringwald M, Rodriguez I, Sakamoto N, Sasaki H, Sato K, Schönbach C, Seya T, Shibata Y, Storch KF, Suzuki H, Toyo-oka K, Wang KH, Weitz C, Whittaker C, Wilming L, Wynshaw-Boris A, Yoshida K, Hasegawa Y, Kawaji H, Kohtsuki S, Hayashizaki Y; RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium. Functional annotation of a full-length mouse cDNA collection. *Nature*. 2001 Feb 8;409(6821):685-90.
- ¹⁹Chhabra ES, Higgs HN. INF2 Is a WASP homology 2 motif-containing formin that severs actin filaments and accelerates both polymerization and depolymerization. *J Biol Chem*. 2006 Sep 8;281(36):26754-67.
- ²⁰Entrez summary for human ortholog.
- ²¹Strausberg RL, Feingold EA, Grouse LH, Derge JG, Klausner RD, Collins FS, Wagner L, Shenmen CM, Schuler GD, Altschul SF, Zeeberg B, Buetow KH, Schaefer CF, Bhat NK, Hopkins RF, Jordan H, Moore T, Max SI, Wang J, Hsieh F, Diatchenko L, Marusina K, Farmer AA, Rubin GM, Hong L, Stapleton M, Soares MB, Bonaldo MF, Casavant TL, Scheetz TE, Brownstein MJ, Usdin TB, Toshiyuki S, Carninci P, Prange C, Raha SS, Loquellano NA, Peters GJ, Abramson RD, Mullahy SJ, Bosak SA, McEwan PJ, McKernan KJ, Malek JA, Gunaratne

PH, Richards S, Worley KC, Hale S, Garcia AM, Gay LJ, Hulyk SW, Villalon DK, Muzny DM, Sodergren EJ, Lu X, Gibbs RA, Fahey J, Helton E, Kettelman M, Madan A, Rodrigues S, Sanchez A, Whiting M, Madan A, Young AC, Shevchenko Y, Bouffard GG, Blakesley RW, Touchman JW, Green ED, Dickson MC, Rodriguez AC, Grimwood J, Schmutz J, Myers RM, Butterfield YS, Krzywinski MI, Skalska U, Smailus DE, Schnerch A, Schein JE, Jones SJ, Marra MA; Mammalian Gene Collection Program Team. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc Natl Acad Sci U S A*. 2002 Dec 24;99(26):16899-903.

²²McCarthy JJ, Meyer J, Moliterno DJ, Newby LK, Rogers WJ, Topol EJ; GenQuest multicenter study. Evidence for substantial effect modification by gender in a large-scale genetic association study of the metabolic syndrome among coronary heart disease patients. *Hum Genet*. 2003 Dec;114(1):87-98.

²³Temel RE, Lee RG, Kelley KL, Davis MA, Shah R, Sawyer JK, Wilson MD, Rudel LL. Intestinal cholesterol absorption is substantially reduced in mice deficient in both ABCA1 and ACAT2. *J Lipid Res*. 2005 Nov;46(11):2423-31.

²⁴Buhman, K.K., M. Accad, S. Novak, R.S. Choi, J.S. Wong, R.L. Hamilton, S. Turley, and R.V.J. Farese. 2000. Resistance to diet-induced hypercholesterolemia and gallstone formation in ACAT2-deficient mice. *Nature Med*. **6**: 1341-1347.

²⁵Repa, J.J., K.K. Buhman, R.B. Farese, Jr., J.M. Dietschy, and S.D. Turley. 2004. ACAT2 deficiency limits cholesterol absorption in the cholesterol-fed mouse: impact on hepatic cholesterol homeostasis. *Hepatology* **40**: 1088-1097.

²⁶Johnson GB, Riggs BL, Platt JL. A genetic basis for the "Adonis" phenotype of low adiposity and strong bones. *FASEB J*. 2004 Aug;18(11):1282-4.

²⁷Ghazalpour A, Doss S, Sheth SS, Ingram-Drake LA, Schadt EE, et al. Genomic analysis of metabolic pathway gene expression in mice. *Genome Biol*. 2005;6:R59.

²⁸Cheng-Chung Tsao, Julie Foley, Sherry J. Coulter, Robert Maronpot, Darryl C. Zeldin, and Joyce A. Goldstein. CYP2C40, a Unique Arachidonic Acid 16-Hydroxylase, Is the Major CYP2C in Murine Intestinal Tract. *Molecular Pharmacology* Vol. 58, Issue 2, 279-287, August 2000

²⁹Ying Hua MA¹, Jia Hua HU¹, Xiao Gang ZHOU¹, Ruo Wang ZENG¹, Zhen Tong MEI², Jian FEI¹ and Li He GUO. Transgenic mice overexpressing γ -aminobutyric acid transporter subtype I develop obesity. *Cell Research* (2000) 10, 303–310.

³⁰Chavey C, Lazennec G, Lagarrigue S, Clapé C, Iankova I, Teyssier J, Annicotte JS, Schmidt J, Matakis C, Yamamoto H, Sanches R, Guma A, Stich V, Vitkova M, Jardin-Watelet B, Renard E, Strieter R, Tuthill A, Hotamisligil GS, Vidal-Puig A, Zorzano A, Langin D, Fajas L. CXC ligand 5 is an adipose-tissue derived factor that links obesity to insulin resistance. *Cell Metab*. (2009) Apr;9(4):339-49.

- ³¹ Chavey C, Fajas L. CXCL5 drives obesity to diabetes, and further. *Aging*. 2009. 1:674-7.
- ³² GeneRIF from Entrez Gene entry.
- ³³ Schneider A, Fischer A, Weber D, von Ahsen O, Scheek S, Krüger C, Rossner M, Klaussner B, Faucheron N, Kammandel B, Goetz B, Herrmann O, Bach A, Schwaninger M. Restriction-mediated differential display (RMDD) identifies pip92 as a pro-apoptotic gene product induced during focal cerebral ischemia. *J Cereb Blood Flow Metab*. 2004. 24:224-36.
- ³⁴ Bradshaw AD, Graves DC, Motamed K, Sage EH. SPARC-null mice exhibit increased adiposity without significant differences in overall body weight. *Proc Natl Acad Sci*. 2003;100(10):6045-50
- ³⁵ Charron Y, Madani R, Nef S, Combepine C, Govin J, Khochbin S, Vassalli JD. Expression of serpinb6 serpins in germ and somatic cells of mouse gonads. *Mol Reprod Dev*. 2006. 73(1):9-19.
- ³⁶ Yuan XD, Wang SJ, Xu YR, Li J, Yang N, Li HF, Gao J. Association of fibrinogen B beta-chain gene polymorphism with factors affecting obesity. *Zhonghua Yi Xue Yi Chuan Xue Za Zhi*. 2010;27(2):198-203.

Lod score plot for chromosomes with multiple eQTLs.

As discussed in the main text, some of the 28 significant SNPs were on the same chromosome and thus a judgment must be made as to whether they are due to the same or different eQTLs. The plots below show the lod scores for the chromosomes with multiple significant SNPs.

Chromosome 1:

There appear to be two separate loci here – one at about 100-110 Mb and one at about 170 Mb.

Chromosome 3:

This seems most likely to be a single locus, but it is not completely obvious. It is possible that the red gene (Lhfp) has two separate peaks. These were grouped as one locus in table 3, but it is noted that Lhfp may have a separate QTL.

chromosome 3, group 422 **chromosome 3, group 518**

Seems to be the same locus.

Chromosome 6:

This one is also ambiguous. The black (Pparg) and red (Mug2) genes clearly share the same eQTL. The green gene (Pde6h) is not obvious – it appears that it may have a separate eQTL. We have classified this as eQTLs in table 3, but this is uncertain.

Chromosome 8:

These are also ambiguous. We judged these as different peaks, but again it is not clear.

Chromosome 12:

This one clearly has two peaks.

Chromosome 14:

This one is also clear: one peak

Chromosome 17:

We judged this as two peaks, although it is not clear cut.

Chromosome 19:

We judged this as two peaks, although it is not clear cut.

Comparing lod score patterns within gene groups.

We wish to determine the extent to which genes in gene groups have shared genetic control. The lod score profiles provide a means of assessing this. The plots below show lod scores for gene groups and chromosomes as shown. These plots show cases in which genes in a group appear to share lod score peaks and therefore are apparently jointly controlled by genetic loci.

We wished to test this more rigorously . That is, we wished to test the hypothesis H_0 : the locations or lod peaks for the genes in groups are independent.

H_1 : the locations of lod peaks for the genes in groups are correlated.

We did this using a randomization approach. First, we found the SNP with the highest lod score for each gene in the group on each chromosome and then calculated the standard deviation in the peak SNP over the genes in the group. We then generated a null distribution as follows using a randomization approach. In each replicate, a peak SNP position is randomly selected for each gene on each chromosome. The standard deviation in this position across genes in the group is calculated for each chromosome. These standard deviations are sorted from lowest to highest. This is repeated 1000 times. At the end of 1000 replicates, we have a matrix where entry (i,j) is the j th lowest standard deviation (over chromosomes) in the i th replicate. We then calculate the lower 5% point for each column. If the k th lowest observed standard deviation is lower than the 5% point for the k th column, the null hypothesis is rejected for that chromosome.

The tables below show the results for gene groups 1-18 (with groups with 1-2 genes excluded). We see that the null hypothesis is rejected for many gene groups on many chromosomes. That is, we see strong evidence of extensive correlation in lod scores among genes in the gene groups.

Gene group															
	1			2			3			5			6		
	5%	obs	chr	5%	obs	chr	5%	obs	chr	5%	obs	chr	5%	obs	chr
1	0.9	5.0	3	7.8	7.5	5	0.9	0	19	6.5	2.8	19	7.9	6.9	10
2	4.2	5.6	10	12.7	7.5	9	4.0	0.4	6	11.4	6.6	2	13.5	7.8	14
3	7.3	6.6	17	16.3	8.1	10	6.9	2.1	14	15.0	16.8	15	16.5	11.5	19
4	9.8	11.2	14	18.9	10.6	14	9.6	4.0	3	17.2	18.0	10	19.0	14.9	12
5	11.7	16.8	19	21.0	10.6	19	12.2	3.0	10	19.3	18.8	18	21.3	18.5	6
6	14.5	18.6	18	23.0	19.5	13	14.5	9.7	9	21.3	19.6	6	22.9	18.9	4
7	16.5	19.8	15	24.5	20.2	11	17.1	12.2	15	23.1	20.2	7	24.4	24.2	15
8	18.5	22.6	7	26.1	20.7	8	19.3	16.2	8	24.8	24.6	17	26.2	25.6	13
9	20.8	24.2	13	29.2	20.8	15	21.0	19.8	11	26.6	25.3	14	27.6	26.1	11
10	23.1	32.1	12	30.6	29.9	18	23.3	21.6	1	28.2	26.4	13	29.2	29.9	18

Gene group															
	7			8			9			10			11		
	5%	obs	chr	5%	obs	chr	5%	obs	chr	5%	obs	chr	5%	obs	chr
1	7.7	3.7	8	8.2	1.1	14	8.2	0.6	3	8.1	0.0	12	0.9	0.0	13
2	13.0	11.4	14	12.6	3.6	19	13.2	5.1	10	13.3	6.6	15	4.1	0.0	17
3	16.4	11.4	10	16.3	13.5	10	16.5	8.1	2	16.8	8.7	14	7.0	1.0	12
4	19.0	13.7	15	18.6	14.5	2	18.9	10.6	15	19.3	18.1	19	9.5	1.4	11
5	21.3	19.0	12	20.9	15.3	6	20.8	14.0	18	21.3	19.3	10	11.9	3.5	2
6	22.8	21.6	18	22.6	17.7	11	22.5	15.6	19	22.8	20.4	18	14.1	5.0	8
7	24.8	26.1	13	24.4	20.9	18	24.4	19.4	14	24.6	30.4	13	16.7	6.5	15
8	26.3	27.4	7	26.3	24.8	15	26.2	29.7	13	26.3	34.1	7	18.7	8.8	3
9	28.0	27.8	19	28.1	27.2	7	27.6	35.1	16	27.8	36.9	16	20.6	17.9	9
10	29.3	32.6	16	29.6	34.4	12	29.1	38.3	17	29.2	40.1	17	23.2	19.1	7

Gene group															
	12			14			15			16			18		
	5%	obs	chr	5%	obs	chr	5%	obs	chr	5%	obs	chr	5%	obs	chr
1	1.0	1.5	16	4.0	0.1	2	8.2	1.1	14	8.0	5.5	14	7.7	0.1	6
2	4.2	6.3	17	8.4	3.3	19	13.4	3.6	19	13.2	8.7	10	13.1	3.1	19
3	7.1	7.9	18	11.5	8.0	10	16.5	13.5	10	16.1	9.5	5	16.3	6.6	15
4	9.5	10.1	14	14.3	9.7	11	18.5	14.5	2	18.9	10.4	19	18.6	7.8	12
5	11.9	13.7	3	16.7	10.6	18	20.9	15.3	6	21.0	14.6	6	20.9	9.1	11
6	13.8	14.4	10	18.8	18.8	15	22.6	17.7	11	23.0	19.7	15	22.9	11.4	10
7	16.8	15.8	15	20.6	20.6	8	24.3	20.9	18	24.6	21.0	8	24.2	20.8	18
8	18.9	16.2	19	22.6	22.6	6	26.0	24.8	15	26.1	28.7	7	25.9	21.4	14
9	21.2	21.6	12	25.0	25.0	13	27.6	27.2	7	27.8	30.5	18	27.4	28.2	17
10	23.6	24.8	11	26.6	26.6	14	29.1	34.4	12	29.4	33.0	12	28.9	30.9	7