Genetics: Published Articles Ahead of Print, published on November 29, 2010 as 10.1534/genetics.110.122796

Title: A model selection approach for expression quantitative trait loci (eQTL) mapping

Authors: Ping Wang[1], John A. Dawson[1], Mark P. Keller[2], Brian S. Yandell[1,3], Nancy A. Thornberry[5], Bei B. Zhang[5], I-Ming Wang[5], Eric E. Schadt[5], Alan D. Attie[2], and C. Kendziorski[4,*]

[1]Department of Statistics, [2]Department of Biochemistry, [3]Department of Horticulture, [4]Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI 53726; [5]Merck, Whitehouse Station, New Jersey, 08889

Running Title: Model selection for eQTL mapping

[*]Corresponding Author:

Christina Kendziorski

Department of Biostatistics and Medical Informatics

University of Wisconsin-Madison

6729 Medical Sciences Center

1300 University Avenue

Madison, WI 53703

Phone: (608) 262-3146

Fax: (608) 265-7916

Email: kendzior@biostat.wisc.edu

Abstract

Identifying the genetic basis of complex traits remains an important and challenging problem with the potential to impact a broad range of biological endeavors. A number of statistical methods are available for mapping quantitative trait loci (QTL), but their application to high throughput phenotypes has been limited as most require user input and interaction. Recently, methods have been developed specifically for expression QTL (eQTL) mapping, but they too are limited in that they do not allow for interactions and QTL of moderate effect. We here propose an automated model-selection based approach that identifies multiple expression quantitative trait loci in experimental populations, allowing for eQTL of moderate effect and interactions. Output can be used to identify groups of transcripts that are likely co-regulated, as demonstrated in a study of diabetes in mouse.

1 Introduction

Many important problems in biology and medicine rely on the accurate identification of the genetic architecture underlying high-throughput phenotypes such as messenger RNA expression. Identifying expression quantitative trait loci (eQTL) and grouping related traits are two primary goals addressed in such endeavors. This manuscript proposes an approach for eQTL mapping and shows how the derived transcript specific genetic signatures can be used to group transcripts that are likely co-regulated.

In the earliest eQTL mapping studies, simple single QTL mapping methods were repeatedly applied to individual expression traits (WILLIAMS *et al.* 2007; KENDZIORSKI and WANG 2006) and that practice continues today. Certainly powerful and effective methods exist providing the flexibility to consider complex genetic models (SEN and CHURCHILL 2001; KAO *et al.* 1999), and they have proven useful in numerous studies. However, the approaches require "fine tuning" (SEN and CHURCHILL 2001) or the choice of thresholds (KAO *et al.* 1999) to resolve multiple linked QTL and identify interactions for a single trait, and as a result applications to expression data are relatively few.

One of the first methods developed specifically for eQTL mapping was proposed by STOREY *et al.* (2005). In that approach, F-statistics are calculated for each marker and trait, and a primary locus is identified for each trait as the one with a maximal F-statistic. A secondary locus is identified as the one having maximal statistic in a second F-test conditional on the first, with permutations used to estimate the posterior probabilities and thresholds for locus-specific and joint linkage. ZOU and ZENG (2009) propose a sequential search for multiple QTL that combines features of Storey's approach with MIM. Both approaches are automated and efficient and therefore useful in eQTL studies. However, the thresholding procedures in identification of primary and

secondary loci may exclude potentially important traits affected by moderate and/or interacting QTL.

The methods discussed thus far all consider trait-specific tests or models, whereas some approaches model all traits (KENDZIORSKI *et al.* 2006; JIA and XU 2007) or groups of traits (CHUN and KELEŞ 2009) at once. With one model for the data, it is possible to account for multiplicities and estimate FDR across transcripts and markers simultaneously. However, the advantage gained is compromised at the level of interacting loci.

In summary, the state-of-the-art QTL mapping methods are sophisticated and quite capable of identifying complicated genetic architecture, but most require that decisions on the class of models to consider, as well as significance thresholds, be made on a case-by-case basis. This clearly limits applications to studies of high-throughput phenotypes such as expression. Many of the challenges have been met to a great extent by the recently proposed methods designed specifically for eQTL mapping. However, these methods are unable to identify eQTL of small or moderate effect, and they do not allow for automated identification of interactions.

We here propose a new multiple QTL mapping approach that has the ability to identify both QTL with large effect and those with small or moderate effect as well as interacting QTL. It is automated and efficient and therefore particularly well suited for eQTL studies. Our approach makes use of the results from a single QTL analysis to reduce the marker search space and thereby reduce the model search space dramatically. The approach is detailed in Section 2.1.

In addition to the multiple eQTL mapping approach, we propose a clustering method which incorporates eQTL mapping results and trait correlations to identify groups of transcripts that likely share similar biological function. An early consideration of this problem is given in EISEN *et al.* (1998) where investigators used hierarchical clustering applied to expression data to identify transcripts with similar function. To date, various clustering algorithms have been proposed in part

to address this same goal (for a comprehensive review, see DO and CHOI (2007)). A particularly powerful and popular approach was proposed by ZHANG and HORVATH (2005). In their work, they describe a module identification approach that uses hierarchical clustering applied to a biologically meaningful distance derived from pairwise correlations between transcripts. When genetic data including genotypes and a genetic map is available in addition to expression data, ideally mapping information can be incorporated to improve the identification of groups of transcripts that are likely co-regulated. To this end, in Section 2.2, we detail an approach that extends ZHANG and HORVATH (2005) to include results from eQTL mapping in the identification of co-expression co-regulation (CECR) modules.

2 Materials and Methods

2.1 A Multiple QTL Identification Approach Allowing for Interactions

Here, we propose a multiple QTL mapping approach that has the ability to identify both QTL with large effect and QTL with small or moderate effect as well as interacting QTL. Motivation for our approach is based on the fact that multiple interacting loci induce marginal effects that can be detected by single QTL mapping methods, as shown for two loci in LAN *et al.* (2001). Given this, the search space for models with first order interactions can be dramatically reduced. Instead of considering interactions between all markers, we focus on markers with relatively high LOD scores, even if those LOD scores are not statistically significant.

2.1.1 Mutiple QTL Mapping Procedure

The approach uses pre-selected markers in a stepwise regression to identify main effects and interactions. Details follow for a single phenotype.

1. Obtain a LOD score profile by applying a single QTL mapping method, such as interval mapping or Haley-Knott regression.

2. Pre-select markers with relatively high LOD scores. Our approach for doing so is provided in the supplement.

3. Perform stepwise regression to obtain a baseline model, one with main effects only. Candidates for main effects in this step are the pre-selected markers and relevant covariates (e.g. sex, age).

4. Perform stepwise regression to obtain the best model with interactions allowed. The potential interactions are between the pre-selected markers or interactive covariates in the baseline model and all pre-selected markers.

In steps 3 and 4, a model selection criterion is needed. Many criteria take the form $-2\log L + k \times c(n)$, where L is the likelihood on n samples given a genetic model with k parameters. For example, $c(n) = 2$ is the classical AIC (AKAIKE 1974); $c(n) = \log(n)$ is the BIC (SCHWARZ 1978). The BIC is used in many studies, but as Broman and Speed (BROMAN and SPEED 2002) point out, its use can result in QTL models with many extraneous variables. ZOU and ZENG (2008) discuss more conservative penalties such as $c(n) = 2\log(n)$ and $c(n) = 3\log(n)$, that we will here refer to as BIC(2) and BIC(3), respectively. The recently proposed penalized LOD score (pLOD) criterion (MANICHAIKUL *et al.* 2009) could also be used.

2.2 A Model Based Clustering Method

In eQTL studies, it is desirable to identify groups of co-regulated traits that share similar biological function. Here we propose a clustering approach designed to accomplish this task. It incorporates both trait correlation and evidence of co-mapping. A measurement to quantify evidence in favor of co-mapping, as measured by the similarity of estimated mapping models, is introduced in 2.2.1.

2.2.1 Similarity between QTL Models

For any pair of models M_1, M_2, defined by the locations of QTL: $M_1 = (q_{11}, q_{12}, \ldots, q_{1n_1})$ and $M_2 = (q_{21}, q_{22}, \ldots, q_{2n_2})$, a similarity measure s should satisfy the following two conditions: i) $s(M_1, M_2) \in [0, 1]$ and ii) $s(M, M) = 1$ for all M.

Assume, without loss of generality, that $n_1 \leq n_2$. Let ϕ_p be a one-to-one mapping from $\{1, \ldots, n_1\}$ to a subset of $\{1, \ldots, n_2\}$ with n_1 elements; there are then $P = \binom{n_2}{n_1} n_1!$ possible mappings. (That is, $\phi_p(i) = \phi_p(j)$ implies $i = j$). We define the model similarity to be

$$s(M_1, M_2) = \frac{2}{n_1 + n_2} \max_{\phi_p} \sum_{i=1}^{n_1} \psi(q_{1i}, q_{2\phi_p(i)}),$$

where ψ is a measure of similarity between two QTL,

$$\psi(q_1, q_2) = \begin{cases} 1 - \frac{r(d)}{r(t)} = \frac{e^{-d/50} - e^{-t/50}}{1 - e^{-t/50}}, & \text{if } q_1, q_2 \text{ on the same chromosome and } d = |q_1 - q_2| \leq m, \\ 0, & \text{otherwise,} \end{cases}$$

m is a parameter set by the user which specifies the genetic distance within which two QTL can be considered similar; $t \geq m$ is a tuning parameter that quantifies the extent of similarity between two QTL within this distance. As t increases, the similarity between any two QTL within the window increases; as t approaches m, the decrease in similarity between two QTL is linear with distance in cM. Supplementary Figure 1 is a plot of similarity between QTL versus distance between QTL in cM when $m = 2.5$cM for various tuning parameters. The choice of m is application dependent.

When small genomic regions are of interest and dense maps and large sample sizes are available, two QTL that are one or two cM apart might not be considered similar. In such a case, m would be chosen to be relatively small compared to situations in which larger regions are of interest with fewer markers and samples. Once m is specified, graphs such as that shown in Supplementary Figure 1 should be used to choose t.

To examine some of the properties of the model similarity defined here, we calculated the similarities among 11 QTL models with 1, 2, and 3 QTL and provide them in Supplementary Tables 1 and 2. As shown there, the similarity measure is a function of QTL proximity between models as well as total number of QTL. Consider for example the similarity calculated between a model M_1 which has a single QTL and a series of nested QTL models M_2, M_5 and M_9, where M_2, M_5, and M_9 each contain a QTL 0.5cM from the QTL in M_1. M_5 and M_9 also contain 1 and 2 additional QTL, respectively. The similarity measure maintains the following ordering: $s(M_1, M_2) > s(M_1, M_5) > s(M_1, M_9)$. This is a desired property since intuitively the similarity between two models should decrease as the number of discrepant loci increases.

2.2.2 Model-Based Clustering Method

A measurement of the adjacency between two traits that incorporates both correlation and mapping information is defined as

$$a_{ij} = |r_{ij}| s_{ij}, \tag{2.1}$$

where r_{ij} is the correlation between traits i and j and

$$s_{ij} = \begin{cases} s(M_i, M_j), \text{ if } s(M_i, M_j) \geq s_0 \\ s_0, \text{otherwise} \end{cases}$$

Instead of directly using $s(M_i, M_j)$ in the definition of a_{ij}, we use s_{ij} so that adjacencies are not zero necessarily for pairs of traits whose model similarities equal 0. Such traits may still be related,

and in this case we allow for the relationship to be directly assessed by correlation. A choice for s_0 is $s_0 = \min\{s(M_i, M_j), s(M_i, M_j) \neq 0\}$.

As in ZHANG and HORVATH (2005), we use average linkage hierarchical clustering coupled with the TOM distance to group traits into modules corresponding to branches of the hierarchical clustering tree (dendrogram). We extend their adjacency measure to the one given in 2.1. Since it accommodates both correlation and co-mapping, we refer to modules constructed using this approach as co-expression co-regulation (CECR) modules.

2.3 Enrichment Test

Given a list of mapping transcripts, it is often of interest to determine whether the transcripts are enriched for any GO (gene ontology) terms in BP (biological process), CC (cellular component), MF (molecular function) categories, or KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways. The hypergeometric test implemented in the R package GOstats was used here for this purpose (R DEVELOPMENT CORE TEAM 2009). The hypergeometric calculation tends to result in small p-values when groups with few transcripts are considered and as a result, it has been suggested that one only consider terms with small p-values and a reasonable number of genes (ten or more) (GENTLEMAN 2005). Unless otherwise stated, we report terms with p-value < 0.001, 10 or more genes on the chip, and 5 or more genes in the list annotated with that term.

3 Data Sets Considered for Evaluation

To assess the proposed methodology we consider many individual traits from the QTL Archive, expression traits collected in a study of diabetes, and simulated data. Details regarding each of these data sources follow.

3.1 QTL Archive Studies

The QTL Archive (http://cgd.jax.org/nav/qtlarchive1.htm) created and maintained by the Jackson Laboratory provides access to raw data and result scripts from various QTL studies using rodent inbred line crosses. There were 31 studies in the QTL Archive as of June 29th, 2008. The mapping method described in Section 2.1 was applied to data from the QTL Archive. BIC was used for model selection, with results evaluated and compared using BIC, BIC(2), and BIC(3).

3.2 Microarray Experiment

The C57BL/6J (B6) and BTBR mice are two inbred mouse populations maintained at the Jackson Laboratory (Bar Harbor, Maine) and often used in studies of type 2 diabetes. When made obese by a leptin mutation, B6 mice are diabetes resistant while the BTBR mice are diabetes susceptible (CLEE *et al.* 2005). In this study, expression profiles were obtained from 499 $F2 - ob/ob$ mice generated from the C57BL/6J (B6) and BTBR founder strains. The profiles probed islet tissue using custom ink-jet microarrays manufactured by Agilent Technologies (Palo Alto, CA). The microarrays consisted of $1,048$ control probes and $39,524$ noncontrol probes. Mouse islets were homogenized and total RNA extracted using Trizol reagent (Invitrogen, CA, USA) according to the manufacturer's protocol. Total RNA was reverse transcribed and labeled with flurochrome. Labeled complementary RNA (cRNA) from each animal was hybridized against a pool of labeled cRNAs constructed from equal aliquots of RNA from all of the animals. All hybridizations were performed in fluor-reversal for 48 hours in a hybridization chamber, washed, and scanned using a confocal laser scanner. Expressions were quantified on the basis of spot intensity relative to background, adjusted for experimental variation between arrays using the average intensity over multiple channels, and fitted to a previously described error model to determine significance

(type I error) (HE *et al.* 2003). Gene expression measures are reported as the ratio of the mean $\log_{10}$ intensity (mlratio). Plasma insulin levels were also measured in each of the 499 mice at approximately 10 weeks of age.

To eliminate the effect of outliers, we performed a normal score transformation based on ranks. In particular, for a trait (insulin level or expression trait) with measurements on n individuals, let R_i be the rank of the measurement for individual i, then the transformed measurement for individual i is $y_i = \Phi^{-1}(R_i/(n+1))$, where Φ^{-1} is the inverse of the standard normal cumulative distribution function. All analyses in this diabetes study are based on the normal scores unless explicitly stated otherwise. Mice were genotyped using the Affymetrix mouse 5K SNP panel (www.affymetrix.com); 1,953 SNPs on 19 autosomes reliably segregated for the founders and were used for QTL mapping.

4 Results

4.1 QTL Archive Studies

There were 31 studies in the QTL Archive as of June 29th, 2008. To be included in our analysis, a study or trait had to satisfy the following conditions: 1) the data set provided in the QTL Archive had to match the description in the paper; 2) the trait to be mapped had to be continuous and suitably handled by the normal model (perhaps following transformation); 3) the markers closest to the identified QTL had to be given explicitly. This results in 24 traits in 11 studies (CLEMENS *et al.* 2000; FARMER *et al.* 2001; MÄHLER *et al.* 2002; LYONS *et al.* 2003a, 2003b; ISHIMORI *et al.* 2004a, 2004b; KORSTANJE *et al.* 2004; LYONS *et al.* 2004a, 2004b; DIPETRILLO *et al.* 2004).

Supplementary Figures 2-12 and Supplementary Tables 3-13 compare the models derived using the proposed approach to those published. As shown in the figures, there is much similarity between models for regions with relatively high LODs. In particular, 67% (63%) of the loci identified in the published models with LODs exceeding 5.0 (4.0) are identified by the proposed approach; 80% (75%) are identified approximately (by markers within 5 cM of the published locus). The published models were also compared to those derived from the proposed approach using standard model selection criteria. Most of the QTL Archive studies derived models using the approach given in SEN and CHURCHILL (2001). As prescribed there, a multiple imputation algorithm is used to fill in missing genotypes. When comparing models derived using the proposed approach to those published, differences due to randomness induced by imputing missing data are not of interest and, as a result, we compare models under two scenarios. The first considers a one time imputation where each model is evaluated on the same set of imputed data; in the second we impute data ten times, evaluate on each set of imputed data, and report the median BIC. BIC^{*} is the BIC obtained with missing genotypes filled in by a one time imputation and BIC^{**} is the median BIC.

Supplementary Table 14 lists BIC^{*} and BIC^{**} corresponding to the published model (superscript 1) and the model identified using the proposed approach (superscript 2). The model complexity, indicated by (# main effects, # interactions), and missing genotype proportions are also given. As suggested by KASS and RAFTERY (1995), we consider two models different if their corresponding BICs differ by more than 10 units. Both BIC^{*} and BIC^{**} suggest that the models identified by the proposed approach are comparable to published models when the amount of missing genotype data is small, and they may be advantageous in some cases. In particular, BIC^{*} (BIC^{**}) associated with the proposed approach is comparable (within ten BIC units) to the BICs derived from published models for 7 of the 16 traits considered when the amount of missing

data is less than 35%. For the 9 traits showing significant difference in BICs, the BICs derived from the proposed approach are smaller. However, when the proportion of missing genotype data exceeds 50%, the proposed approach performs rather poorly, showing comparable BICs in some cases and much larger BICs in others.

As shown in Figure 1, the same result holds generally when BIC(2) and BIC(3) are used. The left panel of Figure 1 shows the adjusted BIC difference between the two models for each trait as $\frac{\text{BIC}^{**,1}-\text{BIC}^{**,2}}{|\text{BIC}^{**,1}+\text{BIC}^{**,2}|/2}$. Similar plots are shown for BIC(2) (middle panel) and BIC(3) (right panel). The traits are ordered (top to bottom) by the proportion of missing genotypes (least to most), as in Supplementary Table 14. Detailed numerical results for the BIC(2) and BIC(3) evaluations are given in Supplementary Tables 15 and 16, respectively.

The results here demonstrate that the models derived using the proposed automated approach largely overlap those found with other methods for regions with relatively high LODs, and most often they show improvement as assessed by the BIC, BIC(2), and BIC(3) when the amount of missing genotype data is relatively small.

4.2 Diabetes Study

This study considers an $F2$ intercross between B57BL/6 (B6) and BTBR mice to study type 2 diabetes. When made obese, B6 mice are resistant to diabetes, whereas BTBR are severely diabetic.

4.2.1 Identification of eQTL and comparison of methods

To identify eQTL and further reveal the genetic architecture underlying expression traits in islet tissue, we applied two mapping approaches, a single QTL mapping approach and the multiple QTL

mapping approach detailed in Section 2.1.

The single QTL mapping approach used here is Haley Knott regression (HALEY and KNOTT 1992), implemented in R/qtl (BROMAN *et al.* 2003). LOD score profiles were obtained at a 2-cM resolution for each trait. For both insulin and the expression traits, sex was included as a main effect and an interactor. A cluster analysis of the 499 mice based on their expression profiles in islet indicated that not only sex, but also the date on which the chips were run had effects on the expression measurements. Therefore, for expression traits, date was also included as a main effect. On each chromosome, the locus with maximum LOD score is claimed as a QTL if the LOD score is greater or equal to 5.0, which controls a genome-wide Type-I error rate at 0.05.

The proposed approach was applied to each trait by first selecting markers with relatively high LOD scores from the Haley Knott regression profiles. The variable search space was reduced dramatically since the numbers of potential marker effects retained from 1,953 markers ranged from 46 to 83. Two stepwise regressions were then performed for model selection, as described in Section 2.1.1, using pLOD as the model selection criterion (MANICHAIKUL *et al.* 2009). As in the single QTL mapping analysis, sex was included as a main effect and a potential interactor for both insulin and the expression traits. For the expression traits, date was also included as a main effect. Table 1 summarizes the complexity of the models for expression traits in islet. In particular, $20,798$ (52.62%) out of the $39,524$ transcripts mapped to at least one QTL. Among the $20,798$ mapping transcripts, 2 or more QTL were identified for 40.44% of the transcripts.

Although it is well known that a multiple QTL mapping analysis is often advantageous over single QTL mapping, a comparison is helpful to determine the particular advantages of the proposed approach. As expected, more eQTL are identified overall using the proposed approach (Figure 2, upper panel). What is perhaps less expected is that the increase is almost entirely due to the identification of additional trans-acting eQTL (eQTL located outside a 5cM window centered

at the physical location of the expression transcript; see the lower panel of Figure 2). In particular, the proposed approach identifies over 92% of the cis-acting QTL (eQTL located within a 5cM window centered on the physical location of the expression transcript) identified by the single QTL mapping approach along with a few others. It also identifies over 80% of the trans-acting QTL identified by a single QTL analysis, but also identifies 50% more trans-acting QTL for most chromosomes. Supplementary Table 17 provides the total counts in detail.

A closer look considers the number of QTL within 5cM windows. Table 2 lists the number of transcripts mapped by each method for several of the hottest windows. Notably, on chromosome 17, the hottest window (one with the most mapping transcripts) from the proposed approach is centered at 17cM while the hottest window from the single QTL analysis is centered at 8.4cM. Interestingly, the transcripts mapped to 17cM through the single QTL analysis did not enrich for any GO BP terms, while those mapped by the proposed approach enriched for GO BP terms mitosis, M phase of mitotic cell cycle, M phase and cell cycle phase with p-value < 0.001. Our group has recently detailed evidence for the role of islet cell-cycle transcripts in diabetes (KELLER *et al.* 2008).

Table 3 shows the results from an enrichment test applied to transcripts mapping to the window on chromosome 6; m.count and s.count are the numbers of genes annotated with the term among the list from the proposed approach and from a single QTL analysis, respectively. For the 29 terms listed, m.count $\geq$ s.count, and for 20 terms, m.p-value $<$ s.p-value, suggesting generally stronger enrichment results for transcripts identified using the proposed approach.

As discussed earlier, one advantage of our proposed approach is the ability to identify interactions, particularly ones that involve moderate main effects. Among the $20,798$ mapping transcripts, sex by marker or marker by marker interactions were identified for $7,985$ (38.39%) transcripts. Among $8,411$ transcripts mapping to 2 or more markers, 797 marker by marker

interactions were identified across 763 transcripts. Figure 3 illustrates the types of interactions identified. The left panel highlights an interaction for which the main effect associated with each interacting term would have been found using the single QTL approach; the middle panel shows a case for which only one of the QTL would have been found; and the right panel a case where neither locus is found significant in a single QTL scan.

4.2.2 Insulin Based Co-Expression Co-Regulation (CECR) Module

When eQTL co-localize with QTL of a clinical trait, one can hypothesize a close relationship exists, such as sharing a regulator (FERRARA *et al.* 2008). The construction of CECR modules has the potential to identify groups of traits that are likely co-regulated, since both the correlation in expression along with mapping information is used. To illustrate, we consider the relationship between insulin and selected expression traits. First, the locations to which insulin maps were identified, where evidence of mapping was quantified using the proposed approach. The model identified for insulin includes 7 QTL and two interactions, one between sex and marker rs3700924 (c17@8.4cM) and the other between sex and marker rs13476801 (c2@91.7cM). The $2,854$ transcripts co-mapping with insulin were then identified as those with at least one locus in common. The pairwise similarities among the $2,855$ traits (insulin and the co-mapping transcripts) were calculated using $m = 2.5$cM and $t = 5$cM, and CECR modules were constructed. Figure 4 shows the resulting modules and the mapping patterns for the traits on the seven chromosomes harboring insulin's QTL. Columns are a series of 5cM non-overlapping bins along the seven chromosomes and each row represents a trait. The much thicker top row highlights the model for insulin, with rows following the top row organized into CECR modules indicated by the colors on the far left. The $(i, j)^{th}$ entry is colored (non-white) if the i^{th} transcript maps to the j^{th} genomic location as assessed by the proposed approach. The color used represents the single QTL LOD score with

LOD scores > 5 shown in black. The top row indicates that insulin maps to 7 locations using the proposed approach, with 2 identified by single QTL mapping.

Enrichment tests were performed to see whether the transcripts in the CECR modules are enriched for any biologically meaningful GO terms or KEGG pathways. The results are listed in Supplementary Table 18. Insulin is in the turquoise module (a module with 540 transcripts) which enriches for innate immune response, a response known to be connected with insulin and diabetes (FERNÁNDEZ-REAL and PICKUP 2008). In contrast, the 540 transcripts most correlated with insulin are enriched only for wound healing, adult behavior, regulation of body fluid levels, and response to virus, none of which is particularly striking. From Figure 4, we see that most transcripts in the turquoise module have QTL near insulin's QTL, rs13483664, at 36.8cM (51Mb) on chromosome 19. SorCS1 is one of them, located on chromosome 19 between 50Mb and 51Mb. In particular, the QTL model for SorCS1 involves rs13483664 and an interaction between sex and rs13483664. CLEE *et al.* (2006) have shown evidence suggesting this gene has broad relevance to the development of type 2 diabetes.

5 Discussion

Many important problems in biology and medicine rely on the accurate identification of QTL contributing to variation in quantitative traits. A number of powerful statistical methods for mapping QTL have proven useful in traditional mapping studies where one or a few quantitative traits are surveyed. Typically, when thousands of phenotypes are available, as in an eQTL mapping study, single QTL mapping methods are repeatedly applied to individual expression traits, effectively sacrificing the identification of refined genetic architecture for efficiency. We here propose an efficient and automated eQTL mapping approach that in part addresses this limitation,

accommodating QTL of small or moderate effect as well as interactions. The output is used to identify CECR modules, groups of transcripts that are likely co-regulated. In practice the approach could and likely should be applied following adjustment for latent variables or population structure such as in LEEK and STOREY (2007) and KANG *et al.* (2008). The effects of doing so were not studied here.

The eQTL mapping approach consists of two stage model selection over a reduced marker search space supported by the fact that interacting QTL induce effects that are detectable marginally. This motivates a first step of identifying markers with relatively large LOD scores following a single QTL scan. Model selection is performed over the selected markers to determine a baseline model of main effects. A second model selection considers possible interactions. As noted, any one of several model selection criteria and procedures for identifying large LODs could be used. The approach results in the identification of a single model per transcript which specifies the QTL affecting that transcript along with their actions and interactions. The model selection methods considered here have a number of advantages, but they do not target error rate control and as a result no statements can be made regarding false discovery rates, for example, either within or across transcripts.

In the analysis of individual traits from the QTL archive, the BIC was used for model selection since many of the published models utilize the BIC to some extent; models were evaluated using the BIC as well as BIC(2) and BIC(3). The results demonstrate that the proposed approach identifies models that largely overlap those found with other methods for loci with large LOD scores, and in most cases they show improvement when there is not a large amount of missing genotype data ($< 35\%$). Here, improvement is assessed by decreased BIC, BIC(2), and BIC(3); and of course in practice it is impossible to know which models better approximate reality. When consideration of one or a few phenotypes is of interest, as in the QTL Archive studies, clearly

multiple approaches and lines of evidence should be considered during model development and selection. However, when high-throughput phenotypes prohibit such a careful and comprehensive evaluation, the automated approach proposed here can be useful.

As demonstrated in the diabetes case study, the output from the proposed approach can be used to identify groups of transcripts (or transcripts and clinical traits) that are likely co-regulated. The so-called CECR module construction extends the definition of module initially proposed by Zhang and Horvath (ZHANG and HORVATH 2005) to accommodate trait specific mapping information, in particular through the specification of a function that defines model similarity. A specific form of model similarity was considered here, but could be modified through the choice of different tuning parameters and/or a different functional form. An investigation of different similarity measures should prove useful in guiding future applications of CECR module construction as well as related efforts that require a measure of distance between two models. In this work, CECR module construction was used to identify groups of transcripts correlated and co-mapping with insulin. The groups generally show stronger enrichment for biological functions, suggesting improvement over using correlation measures alone.

Acknowledgments: This work was supported in part by NIGMS 76274, NIDDK 66369, NIDDK 58037, and NIGMS training grant 74904.

References

AKAIKE, H., 1974 A new look at statistical model identification. IEEE Transactions on Automatic Control **AC-19:** 716–723.

BROMAN, K. W. and T. P. SPEED, 2002 A model selection approach for the identification of quantitative trait loci in experimental crosses. J. Roy. Stat. Soc. B **64:** 641–656.

BROMAN, K. W., H. WU, S. SEN, and G. A. CHURCHILL, 2003 R/qtl: QLT mapping in experimental crosses. Bioinformatics **19:** 889–890.

CHUN, H. and S. KELEŞ, 2009 Expression quantitative trait loci mapping with multivariate sparse partial least squares regression. Genetics **182:** 79–90.

CLEE, S. M., S. T. NADLER, and A. D. ATTIE, 2005 Genetic and genomic studies in the BTBR ob/ob Model of type 2 diabetes. Am. J. Ther. **12:** 491–498.

CLEE, S. M., B. S. YANDELL, K. M. SCHUELER, M. E. RABAGLIA, and O. C. RICHARDS *et al.*, 2006 Positional cloning of a type 2 diabetes quanitative trait locus. Nature Genetics **38:** 688–693.

CLEMENS, K. E., G. CHURCHILL, N. BHATT, K. RICHARDSON, and F. P. NOONAN, 2000 Genetic control of susceptibility to UV-induced immunosuppression by interacting quantitative trait loci. Genes and Immunity **1(4):** 251–259.

DIPETRILLO, K., S.-W. TSAIH, S. SHEEHAN, C. JOHNS, and P. KELMENSON *et al.*, 2004 Genetic analysis of blood pressure in C3H/HeJ and SWR/J mice. Physiological Genomics **17(2):** 215–220.

DO, J. H. and D. K. CHOI, 2007 Clustering approaches to identifying gene expression patterns from DNA microarray data. Mol. Cells **25(2):** 279–288.

EISEN, M. B., P. T. SPELLMAN, P. O. BROWN, and D. BOTSTEIN, 1998 Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. U. S. A. **95:** 14863–14868.

FARMER, M. A., J. P. SUNDBERG, I. J. BRISTOL, G. A. CHURCHILL, and R. LI *et al.*, 2001 A major quantitative trait locus on chromosome 3 controls colitis severity in IL-10-deficient mice. PNAS **98(24):** 13820–13825.

FERNÁNDEZ-REAL, J. M. and J. C. PICKUP, 2008 Innate immunity, insulin resistance and type 2 diabetes. Trends Endocrinol Metab. **19(1):** 10–16.

FERRARA, C. T., P. WANG, E. C. NETO, R. D. STEVENS, and J. R. BAIN *et al.*, 2008 Genetic networks of liver metabolism revealed by integration of metabolic and transcriptional profiling. PLoS Genet. **4(3):** e1000034.

GENTLEMAN, R., 2005 Using GO for Statistical Analyses. Bioconductor Vignettes.

HALEY, C. S. and S. A. KNOTT, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity **69:** 315–324.

HE, Y. D., H. DAI, E. E. SCHADT, G. CAVET, and S. W. EDWARDS *et al.*, 2003 Microarray standard data set and figures of merit for comparing data processing methods and experiment designs. Bioinformatics **19:** 956–965.

ISHIMORI, N., R. LI, P. M. KELMENSON, R. KORSTANJE, and K. A. WALSH *et al.*, 2004a Quantitative Trait Loci Analysis for Plasma HDL-Cholesterol Concentrations and Atherosclerosis Susceptibility Between Inbred Mouse Strains C57BL/6J and 129S1/SvImJ. Arterioscler. Thromb. Vasc. Biol. **24:** 161–166.

ISHIMORI, N., R. LI, P. M. KELMENSON, R. KORSTANJE, and K. A. WALSH *et al.*, 2004b Quantitative trait loci that determine plasma lipids and obesity in C57BL/6J and 129S1/SvImJ

inbred mice. Journal Lipid Research **45:** 1624–1632.

JIA, Z. and S. XU, 2007 Mapping quantitative trait loci for expression abundance. Genetics **176:** 611–623.

KANG, H. M., C. YE, and E. ESKIN, 2008 Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. Genetics **180:** 1909–1925.

KAO, C. H., Z.-B. ZENG, and R. D. TEASDALE, 1999 Multiple interval mapping for quantitative trait loci. Genetics **152:** 1203–1216.

KASS, R. E. and A. E. RAFTERY, 1995 Bayes Factors. Journal of the American Statistical Association **90:** 773–795.

KELLER, M. P., Y. J. CHOI, P. WANG, D. B. DAVIS, and M. E. RABAGLIA *et al.*, 2008 A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility. Genome Research **18:** 706–716.

KENDZIORSKI, C., M. CHEN, M. YUAN, H. LAN, and A. D. ATTIE, 2006 Statistical methods for expression quantitative trait loci (eQTL) Mapping. Biometrics **62:** 19–27.

KENDZIORSKI, C. and P. WANG, 2006 On statistical methods for expression quantitative trait loci mapping. Mammalian Genome **17(6):** 509–517.

KORSTANJE, R., R. LI, T. HOWARD, P. KELMENSON, and J. MARSHALL *et al*, 2004 Influence of sex and diet on quantitative trait loci for HDL cholesterol levels in an SM/J by NZB/BlNJ intercross population. Journal of Lipid Research **45:** 881–888.

LAN, H., C. M. KENDZIORSKI, L. A. SHEPEL, J. D. HAAG, and M. A. NEWTON *et al.*, 2001 Genetic loci controlling breast cancer susceptibility in the Wistar-Kyoto rat. Genetics **157:** 331–339.

LEEK, J. T. and J. D. STOREY, 2007 Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genetics **3(9):** 1724–1735.

LYONS, M. A., R. KORSTANJE, R. LI, K. A. WALSH, and G. A. CHURCHILL *et al.*, 2004a Genetic contributors to lipoprotein cholesterol levels in an intercross of 129S1/SvImJ and RIIIS/J inbred mice. Physiological Genomics **17:** 114–121.

LYONS, M. A., H. WITTENBURG, R. LI, K. A. WALSH, and G. A. CHURCHILL *et al.*, 2003a Quantitative trait loci that determine lipoprotein cholesterol levels in DBA/2J and CAST/Ei inbred mice. Journal of Lipid Research **44:** 953–967.

MÄHLER, M., C. MOST, S. SCHMIDTKE, J. P. SUNDBERG, and R. LI *et al.*, 2002 Genetics of Colitis Susceptibility in IL-10-Deficient Mice: Backcross versus F2 Results Contrasted by Principal Component Analysis. Genomics **80(3):** 274–282.

MANICHAIKUL, A., J. Y. MOON, S. SEN, B. S. YANDELL, and K. W. BROMAN, 2009 A model selection approach for the identification of quantitative trait loci in experimental crosses, allowing epistasis. Genetics **181:** 1077–1086.

R DEVELOPMENT CORE TEAM, 2009 R: A language and environment for statistical computing.

SCHWARZ, G., 1978 Estimating the dimension of a model. The Annals of Statistics **6:** 461–464.

SEN, S. and G. A. CHURCHILL, 2001 A statistical framework for quantitative trait. Genetics **159:** 371–387.

STOREY, J. D., A. J. M, and K. L, 2005 Multiple locus linkage analysis of genomewide expression in yeast. PLoS Biology **3(8):** e267.

WILLIAMS, R. B., E. K. CHAN, M. J. COWLEY, and P. F. LITTLE, 2007 The influence of genetic variation on gene expression. Genome Research **17:** 1707–1716.

ZHANG, B. and S. HORVATH, 2005 A general framework for weighted gene co-expression network analysis. Stat. Appl. Genet. Mol. Biol. **4:** Article 17.

ZOU, W. and Z.-B. ZENG, 2008 Statistical methods for mapping multiple QTL. International Journal of Plant Genomics **Article ID:** 286561.

ZOU, W. and Z.-B. ZENG, 2009 Multiple interval mapping for gene expression QTL analysis. Genetica **137:** 125–134.

Table 1: The number of transcripts having 1,...,7 and more than 7 main effects. The totals are given as percentages of the $39,524$ transcripts (percentage 1) and the $20,798$ mapping transcripts (percentage 2).

# of QTL	1	2	3	4	5	6	7	>7	total
# of transcripts	12387	4877	1960	849	407	184	84	50	20798
percentage1	31.34	12.34	4.96	2.15	1.03	0.47	0.21	0.13	52.62
percentage2	59.56	23.45	9.42	4.08	1.96	0.88	0.4	0.24	100

Table 2: The number of transcripts mapping in 5cM windows identified by the single and multiple QTL analysis (n.mapping.s and n.mapping.m, respectively). Position of the window center is shown in cM.

chr	Pos (cM)	n.mapping.s	n.mapping.m
6	108.1	1978	2373
2	96.0	1578	1678
2	73.0	959	1592
2	89.2	796	984
7	4.0	495	890
12	8.0	402	843
17	8.4	621	778

Table 3: Results from an enrichment test applied to transcripts mapping to the 5cM window centered at 108.1 cM on chromosome 6. Listed are terms with size ≥ 10 and p-value ≤ 0.001 on either list from the single (s) and multiple (m) QTL analysis.

set	term	size	s.count	s.p-value	m.count	m.p-value
GOBP	multicellular organismal process	3774	236	9.56e-07	292	5.07e-08
GOBP	multicellular organismal development	2087	137	5.04e-05	166	2.86e-05
GOBP	cell adhesion	585	44	0.0021	59	4.42e-05
GOBP	biological adhesion	585	44	0.0021	59	4.42e-05
GOBP	phosphate transport	76	8	0.029009	14	0.000135
GOBP	organ development	1319	89	0.000554	108	0.000317
GOBP	tube development	190	22	0.000118	24	0.000384
GOBP	system development	1612	103	0.001324	126	0.000651
GOBP	proteolysis	605	47	0.000763	55	0.001063
GOBP	tube morphogenesis	134	16	0.000692	17	0.002481
GOBP	embryonic development	437	37	0.000583	40	0.004356
GOCC	proteinaceous extracellular matrix	278	30	4.54e-05	42	3.87e-08
GOCC	extracellular matrix	282	30	5.92e-05	42	5.86e-08
GOCC	extracellular region	2483	164	2.44e-05	210	1.03e-07
GOCC	collagen	37	6	0.008744	13	1.27e-07
GOCC	extracellular matrix part	93	10	0.015876	20	5.82e-07
GOCC	extracellular region part	2037	133	0.000289	172	2.34e-06
GOCC	extracellular space	1919	118	0.005395	156	5.94e-05
GOCC	intrinsic to plasma membrane	648	51	0.000605	60	0.000695
GOMF	extracellular matrix structural constituent conferring tensile strength	29	6	0.00263	13	4.09e-09
GOMF	extracellular matrix structural constituent	59	8	0.008438	16	2.99e-07
GOMF	peptidase activity	625	50	0.000599	59	0.000545
GOMF	metalloendopeptidase activity	105	16	5.87e-05	16	0.000608
GOMF	cytokine activity	214	18	0.020571	26	0.000619
GOMF	transmembrane receptor activity	1891	124	0.000615	147	0.000961
GOMF	endopeptidase activity	408	37	0.000307	41	0.001155
GOMF	carbohydrate binding	259	26	0.000526	28	0.002367
KEGG	Cell Communication	125	16	0.000246	23	1.72e-06
KEGG	ECM-receptor interaction	84	11	0.001933	16	4.47e-05

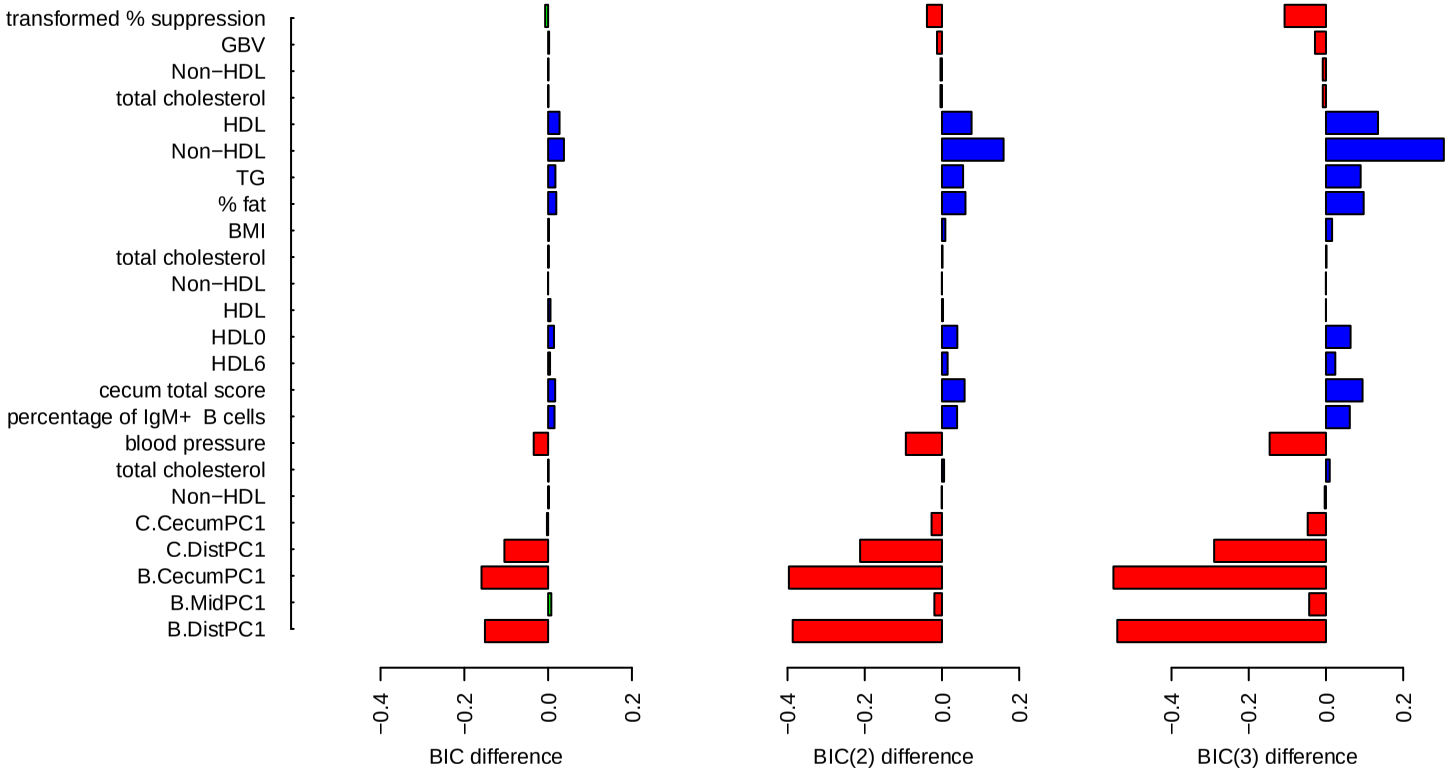

Figure 1: Adjusted BIC difference for QTL Archive studies. Positive (negative) absolute differences equal to or exceeding 10 units are highlighted in blue (red); absolute differences smaller than 10 units are highlighted in green.

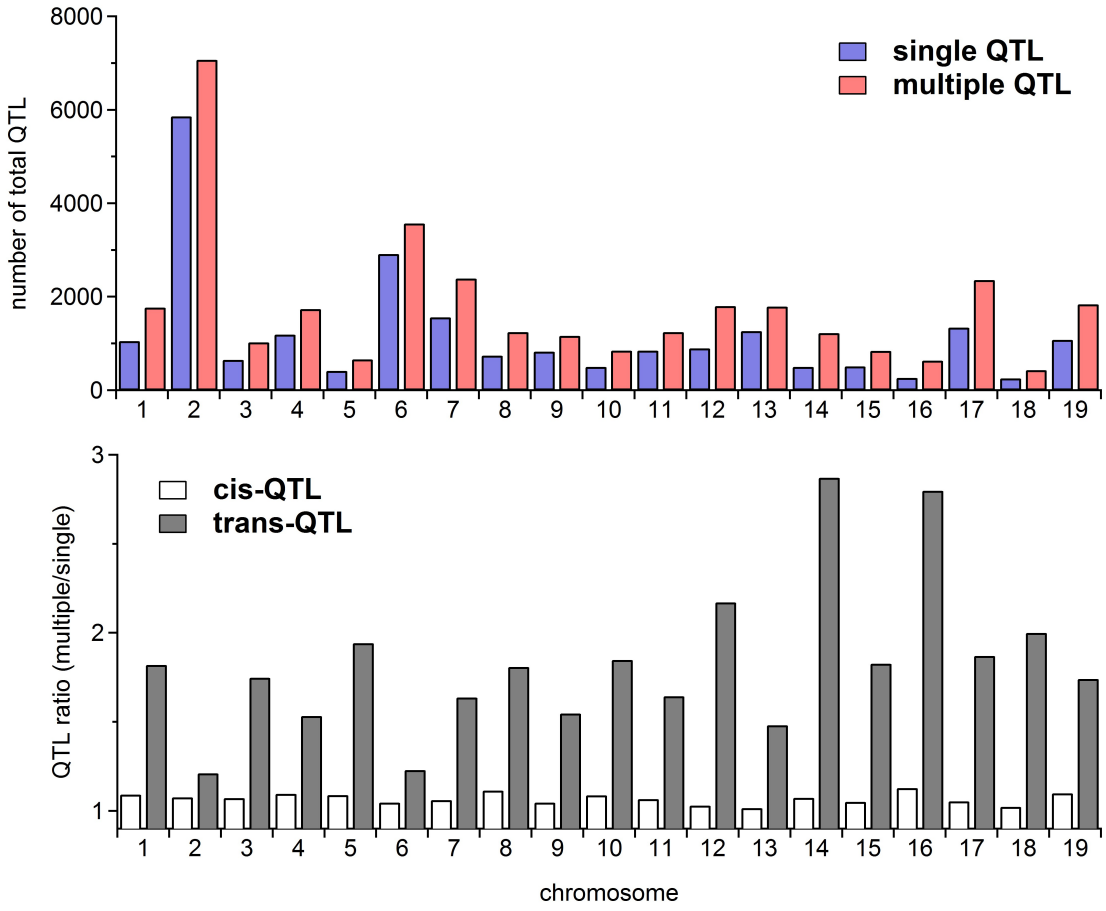

Figure 2: The upper panel shows the total number of QTL (cis-QTL + trans-QTL) identified by the single and multiple QTL mapping approaches. The lower panel compares the number of cis-QTL identified by the multiple QTL mapping approach to that identified using single QTL mapping (white bars), and similarly for trans-QTL (grey bars).

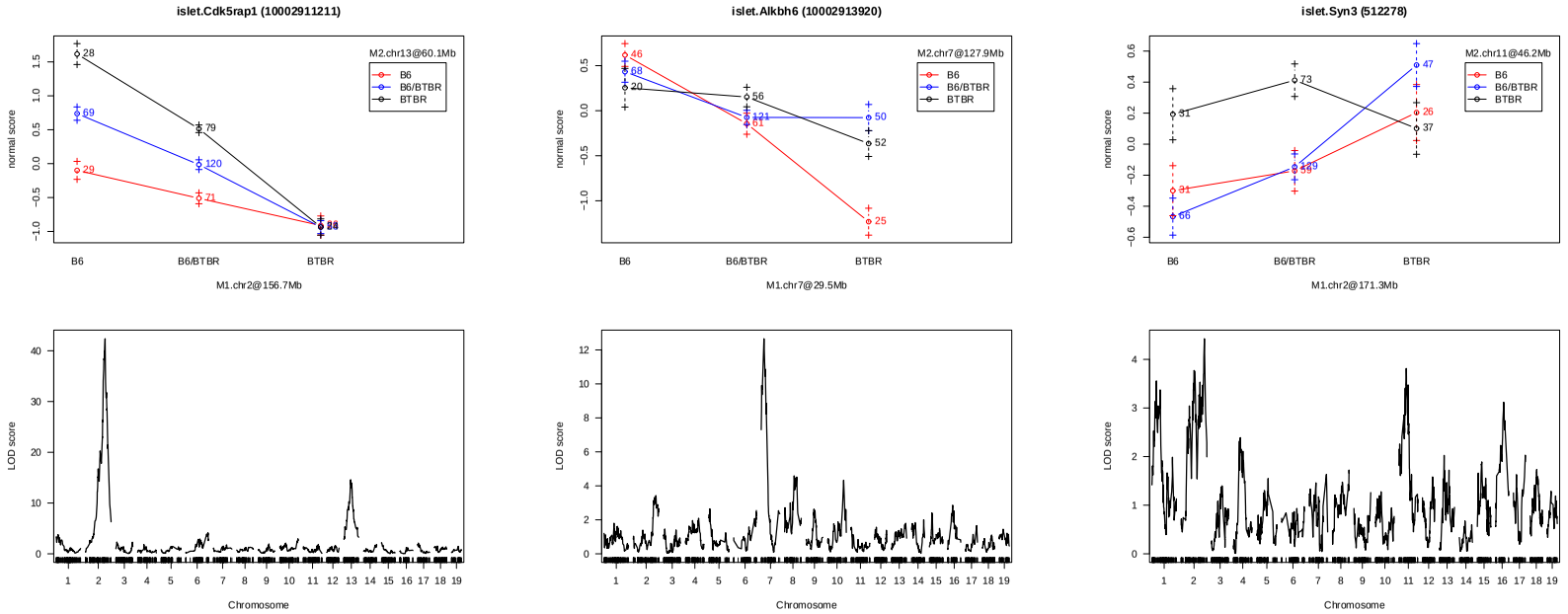

Figure 3: Interaction plots and single QTL LOD profiles for three expression traits. The LOD score for Cdk5rap1 (left panel) at rs4223605 (M1.chr2 at 156.7Mb) is 42.37 and at rs13481837 (M2.chr13 at 60.1Mb) is 14.6; the LOD score for Alkbh6 (middle panel) at rs4226520 (M1.chr7 at 29.5Mb) is 12.66 and at rs13479518 (M2.chr7 at 127.9Mb) is 0.64; and the LOD score for Syn3 (right panel) at rs13476918 (M1.chr2 at 171.3Mb) is 4.43 and at rs6365385 (M2.chr11 at 46.2Mb) is 3.81.

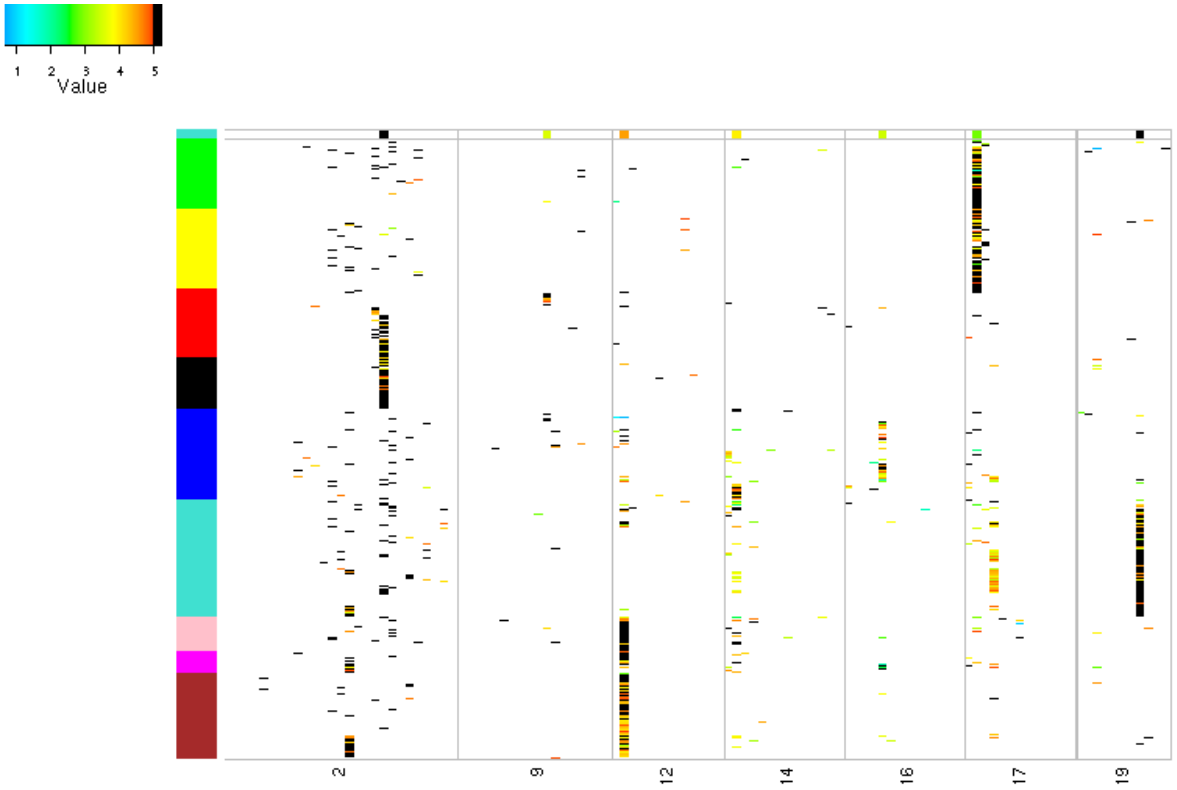

Figure 4: The mapping patterns for insulin and the $2,854$ co-mapping transcripts. Columns are a series of 5cM non-overlapping bins across 7 chromosomes harboring locations to which insulin maps. Shown are $2,855$ rows. The first represents insulin and is extra thick so that the locations to which insulin maps can be easily seen. There are $2,854$ rows following, one for each transcript, with row ordering determined by the CECR module construction. The color bar at the left represents the CECR modules. Insulin is in the turquoise module. Bins containing QTL are colored (non-white) with the color representing the magnitude of the LOD score obtained from a single QTL analysis (LODs > 5 are colored black). In the rare event that a bin contains more than one QTL, the color corresponds to the maximum LOD score.