Published in final edited form as:

J Am Stat Assoc. 2015 ; 110(512): 1350–1361. doi:10.1080/01621459.2015.1049746.

Bayesian Partition Models for Identifying Expression Quantitative Trait Loci

Bo Jiang and

Harvard University, Cambridge, MA 02138

Jun S. Liu [Professor of Statistics]

Department of Statistics, Harvard University, Cambridge, MA 02138

Abstract

Expression quantitative trait loci (eQTLs) are genomic locations associated with changes of expression levels of certain genes. By assaying gene expressions and genetic variations simultaneously on a genome-wide scale, scientists wish to discover genomic loci responsible for expression variations of a set of genes. The task can be viewed as a multivariate regression problem with variable selection on both responses (gene expression) and covariates (genetic variations), including also multi-way interactions among covariates. Instead of learning a predictive model of quantitative trait given combinations of genetic markers, we adopt an inverse modeling perspective to model the distribution of genetic markers conditional on gene expression traits. A particular strength of our method is its ability to detect interactive effects of genetic variations with high power even when their marginal effects are weak, addressing a key weakness of many existing eQTL mapping methods. Furthermore, we introduce a hierarchical model to capture the dependence structure among correlated genes. Through simulation studies and a real data example in yeast, we demonstrate how our Bayesian hierarchical partition model achieves a significantly improved power in detecting eQTLs compared to existing methods.

Keywords

Bayesian Variable Selection; Dirichlet Process; Expression Quantitative Trait Loci; Hierarchical Model; Interaction Detection

1 Introduction

The most common type of genetic variation among living organisms is called Single Nucleotide Polymorphism (SNP). Each SNP represents a single nucleotide position in the genome that has been observed to have different nucleotide types among members of one species. Current practices for human genetics usually require that the least frequent type (minor allele) occurs in at least 1% of the population. On average SNPs occur once in every 300 nucleotides in the human genome, and they occur much more frequently in lower organisms such as the budding yeast. Expression quantitative trait loci (eQTLs) refer to

genomic loci associated with changes of expression levels of certain genes. By assaying gene expression and genetic variation (*e.g.*, SNPs and/or copy number variations (CNVs)) simultaneously in segregating populations, scientists wish to correlate variations in the gene expression with genomic sequence variations. In such cases we say that a gene's expression is linked to or maps to the corresponding genetic loci, and thus likely regulated by genomic regions surrounding those loci. One justification for studying genetics of gene expression is that transcript abundance may act as an intermediate phenotype between genomic sequence variation and more complex whole-body phenotypes. Results from eQTL studies have been used for identifying hot spots (Brem *et al.*, 2002; Schadt *et al.*, 2003; Morley *et al.*, 2004; Bystrykh *et al.*, 2005; Chesler *et al.*, 2005; Hubner *et al.*, 2005; Lan *et al.*, 2006), constructing causal networks (Zhu *et al.*, 2004; Bing and Hoeschele, 2005; Chesler *et al.*, 2005; Li *et al.*, 2005; Schadt *et al.*, 2005; Zhu *et al.*, 2008), prioritizing lists of candidate genes for clinical traits (Bystrykh *et al.*, 2005; Hubner *et al.*, 2005; Schadt *et al.*, 2005), and elucidating subclasses of clinical phenotypes (Schadt *et al.*, 2003; Bystrykh *et al.*, 2005).

Traditional eQTL studies are based on linear regression models (Lander and Botstein, 1989) in which each trait variable is regressed against each marker variable. The *p*-value of the regression slope is reported as a measure of significance for association. In the context of multiple traits and markers, procedures such as false discovery rate (FDR) controls (Benjamini and Hochberg, 1995; Storey and Tibshirani, 2003) can be used to control family-wise error rates. Despite the success of regression approaches in detecting single eQTLs, a number of challenging problems remain. First, these methods can not easily discover *epistasis effect*, *i.e.*, the joint effect of multiple markers. Storey *et al.* (2005) developed a step-wise regression method to search for pairs of markers. This procedure, however, tends to miss eQTL pairs with small marginal effects but a strong interaction effect. Second, there are often strong correlations among expression levels for groups of genes (called *gene modules*), partially reflecting co-regulation of genes in biological pathways that may respond to common genetic loci and environmental perturbations (Schadt *et al.*, 2003; Yvert *et al.*, 2003; Chen *et al.*, 2008; Schadt *et al.*, 2008; Zhu *et al.*, 2008). Previous findings of eQTL *hot spots*, *i.e.*, loci affecting a larger number of expression traits, and their biological implications further enhance this notion and highlight the biological importance of finding such *pleiotropic effects*.

Mapping genetic loci for multiple traits simultaneously has also been shown to be more powerful than mapping single traits at a time (Jiang and Zeng, 1995). Although for a known small set of correlated traits, one can conduct QTL mapping for a few principal components (Mangin *et al.*, 1998), this type of methods becomes ineffective when the set size is moderately large or one has to enumerate all possible subsets. An alternative approach is to identify subsets of genes by a clustering method in the first stage, and then fit mixture models to clusters of genes (Kendzioriski *et al.*, 2006) or linear regression by treating genes as multivariate responses (Chun and Kele, 2009). The eQTL mapping then depends on whether the clustering method can find the right number of clusters and the right gene partitions.

The problem of searching for eQTLs can be viewed as a variable selection problem, selecting on both predictors (genotypes of SNPs) and responses (gene expression), including

also multi-way interactions among the predictors. Variable selection in regression modeling is a long-standing problem in statistics, especially in analyzing high-dimensional and high-throughput data. Traditional variable selection methods, from which most of the aforementioned methods are derived, focus on the forward modeling perspective, *i.e.*, predictive modeling for the conditional distribution of response(s) Y given predictors X . Our goal here is to detect nontrivial joint effects of subsets of predictors on the response vector. Traditional approaches are therefore rather cumbersome to use and sensitive to distributional assumptions since it needs to (a) specify how multiple predictors interact (*e.g.*, a multiplicative effect), and (b) include all possible interaction terms as candidates. As the number of possible genotype combinations grows exponentially with the number of SNPs under consideration, it is very likely that some genotype combinations contain very few or even no observations, and regression-based methods such as analysis of variance (ANOVA) have only limited power in such situations.

In contrast to the forward regression formulation, Zhang and Liu (2007) introduced the Bayesian epistasis association mapping (BEAM) model to detect epistatic interactions in genome-wide case-control studies, where response Y is a binary variable indicating disease status. The BEAM model can be viewed as a generalization of the naïve Bayes (NB) model, which models $\Pr(X|Y)$ instead of $\Pr(Y|X)$. Motivated by the success of BEAM, Zhang *et al.* (2010) developed a Bayesian partition (BP) model for eQTL studies based on a joint model of gene expression and SNPs. More specifically, correlated expression traits Y and their associated set of markers X are treated as a *module* in the BP model and a latent *individual type* variable T is introduced to decouple X and Y by modeling $\Pr(X|T)$ and $\Pr(Y|T)$ separately. A Markov Chain Monte Carlo (MCMC) algorithm (Liu, 2008) was used to search for the module genes and their linked markers. Compared with regression-based approaches, the Bayesian partition model offers a greater flexibility in modeling and searching for epistatic effects.

The BP model in Zhang *et al.* (2010) has several limitations in its flexibility and scalability due to its restrictive model assumptions and high computational costs. First, it only allows positively correlated genes to be selected into the same module and cannot capture complex gene expression patterns in a module. Second, the individual types in the original BP model are determined using an *ad hoc* approach, violating MCMC sampling rules. Third, the joint distribution of all the associated markers in a module is described by a saturated model with an exponentially growing complexity, which decreases the model's power in detecting multi-SNP associations, especially for markers that are only marginally associated with a module. Moreover, to account for linkage disequilibrium (LD) among adjacent markers, the original BP model imposed a mutually exclusive condition on marker pairs with correlations exceeding a certain threshold, which is somewhat artificial. Last but not least, the original MCMC algorithm converges slowly because it needs to iterate through a large number of intermediate parameters. Although a parallel tempering scheme had been employed to help with the mixing of the chain, it still requires intensive computational resources.

In this article, we propose and implement the second-generation Bayesian partition model (henceforth, BP2 model) and its associated efficient MCMC algorithm to address limitations of the previous BP model. Under a Bayesian framework with latent individual types, BP2

model uses additional latent variables to partition genes into positively correlated gene clusters and aggregate multiple gene clusters into a module. Clustering of genes makes the computation faster and alleviates the dominance of the gene expression clustering effect in module determination. The aggregation of multiple gene clusters into a module allows the model to capture the complex dependence structure among gene expression such as negative co-expression. The BP2 model introduces a flexible Chinese restaurant process to model individual types and draws posterior samples of individual types within a principled Gibbs sampling framework. The BP2 model also divides SNPs in a module into independent marker groups modeled separately by saturated multinomial models, which increases its ability in detecting weak marginal effects. The BP2 model further improves upon the BP model by modeling the block structure of LD and selecting SNPs within blocks that are associated with gene expression, either individually or interactively with other SNPs. By collapsing (integrating out) intermediate parameters in the hierarchical model, the convergence of the associated MCMC algorithm has also been significantly accelerated.

The rest of this paper is organized as follows: we start in Section 2 with an overview of BP2 model and then describe different components of the partition model in details. Simulation studies that compare the BP2 with regression-based methods and the previous BP method are presented in Section 4. In Section 5, we illustrate our method on a yeast eQTL data set. We conclude the paper with a short discussion.

2 Bayesian partition model for eQTLs

Let Y_j be the quantile normalized and standardized expression level of gene $j \in \{1, 2, \dots, q\}$, and let X_k ($k \in \{1, \dots, p\}$) be a categorical variable with support $\{1, \dots, V\}$, representing the genotype of a SNP. Throughout this section, we use boldface fonts to denote realizations of random vectors, and use $\Pr(\mathbf{x}_{\mathcal{S}}|\mathbf{y}_{\mathcal{R}})$ as a shorthand notation for the conditional probability of observing $\{X_{i,k} = x_{i,k}\}_{k \in \mathcal{S}}$ given $\{Y_{i,j} = y_{i,j}\}_{j \in \mathcal{R}}$ ($i = 1, \dots, n$), that is,

$$\Pr(\mathbf{x}_{\mathcal{S}}|\mathbf{y}_{\mathcal{R}}) := \prod_{i=1}^n \Pr(\{X_{i,k} = x_{i,k}\}_{k \in \mathcal{S}} | \{Y_{i,j} = y_{i,j}\}_{j \in \mathcal{R}}),$$

where $\mathcal{S}$ and $\mathcal{R}$ are some index sets of random variables X_k and Y_j .

We define an eQTL “module” as a set of gene expression traits and a set of SNPs such that the variation of the gene expression traits is associated with the genotype combination of the SNPs. This association between multiple genes and multiple SNPs is characterized by a latent variable T , which represents a partition of all the individuals and is termed as “individual type” henceforth. A realization of T partitions all individuals into subgroups of the “same-type” ones. Gene expression traits and SNPs are conditionally independent given the individual type. The goal of the Bayesian partition method is to simultaneously assign gene expression traits and SNPs into modules. We start by giving an overview of partition model for eQTL modules before diving into individual model components in details.

2.1 Overview of partition model for eQTL modules

The Bayesian partition model includes D modules (the choice of D will be discussed in Section 3.2) with each module consisting of one or more clusters of genes and a set of SNP candidates for quantitative trait loci (QTLs). Gene clusters are building blocks of modules. Genes are divided into clusters with positively correlated expression levels. We use C_j to denote the cluster membership of gene j ($j = 1, \dots, q$), and define index set $\mathcal{G}_c = \{j: C_j = c\}$ ($c = 1, \dots, K$ and K is assumed to fixed here) and their observed expression values $\mathbf{y}_{\mathcal{G}_c} = \{y_{ij}: j \in \mathcal{G}_c, i = 1, \dots, n\}$. The set of genes that do not belong to any cluster is denoted as $\mathcal{G}_0 = \{j: C_j = 0\}$, and we assume that their expression values (after quantile normalization) follow independent Gaussian distributions. Each gene cluster is assigned to at most one module and clusters within the same module have correlated expression patterns (either positively or negatively). We use J_c to denote the module membership of cluster c , which equals to d if the gene cluster belongs to the eQTL module indexed by d and 0 if the gene cluster does not belong to any module. Note that although genes from two different clusters in the same module share the same individual type partition, they can be negatively correlated with each other.

SNPs are modeled separately for each module and different modules can share the same SNP (see Supplementary Materials for further discussions on this assumption). In other words, every module has its own “copy” of the entire genome, from which we want to select a subset of SNPs that are associated with (or determine) the individual type, which is then associated with the expression pattern of gene clusters. We define the association indicator $I_{k,d}$ for SNP k ($k = 1, \dots, p$) and module d ($d = 1, \dots, D$), where $I_{k,d} = 1$ if the marker is associated with the module indexed by d and $I_{k,d} = 0$ otherwise. We use $\mathcal{A}_d = \{k: I_{k,d} = 1\}$ to denote the set of associated SNPs, i.e. QTLs, and $\Pr(\mathbf{x}_{\mathcal{A}_d^c} | \mathbf{x}_{\mathcal{A}_d})$ to denote the conditional distribution of all other SNPs given the set of QTLs in module d .

The association between gene clusters and QTLs in a module is characterized by the common latent individual type partition. As illustrated in Figure 1, conditioning on individual types $\mathbf{t}_d = \{t_{d,i}\}_{i=1}^n$ for module d , each gene cluster in module d , $\mathbf{y}_{\mathcal{G}_c}$ given $J_c = d$ (i.e., cluster c is assigned to module d) and the set of QTLs, $\mathbf{x}_{\mathcal{A}_d}$ are modeled independently, which are denoted as $\Pr(\mathbf{x}_{\mathcal{A}_d} | \mathbf{t}_d)$ and $\Pr(\mathbf{y}_{\mathcal{G}_c} | \mathbf{t}_d)$, respectively. Furthermore, we assume that the individual type T_d follows a Chinese restaurant process *a priori* and the joint prior probability of observing $\mathbf{t}_d = \{t_{d,i}\}_{i=1}^n$ can be written as

$$\Pr(\mathbf{t}_d) = \left[\frac{\omega_0^{|T_d|} \prod_{t=1}^{|T_d|} (n_t - 1)!}{\omega_0(1+\omega_0) \dots (n-1+\omega_0)} \right], \quad (1)$$

where n_t is the number of observations with individual type t , $|T_d|$ is the number of distinct individual types in $\mathbf{t}_d$ and ω_0 is a pre-specified concentration parameter.

Three sets of parameters in the partition model are of interest to us: SNP association indicators $I = \{I_{k,d}\}_{k=1, \dots, p, d=1, \dots, D}$ with each $I_{k,d} \in \{0, 1\}$, gene cluster indicators $C = \{C_j\}_{j=1, \dots, q}$

with each $C_j \in \{1, \dots, K\}$, and module membership of clusters $J = \{J_c\}_{1 \leq c \leq K}$ with each $J_c \in \{1, \dots, D\}$. Let η_C , η_J and η_I be the prior probabilities of adding a gene into a cluster, adding a cluster to a module and adding a SNP to a module, respectively. Our prior on parameters of interest is given by

$$\Pr(I, J, C) \propto \left(\frac{\eta_I}{1 - \eta_I}\right)^{N_I} \left(\frac{\eta_J}{1 - \eta_J}\right)^{N_J} \left(\frac{\eta_C}{1 - \eta_C}\right)^{N_C},$$

where $N_C = \sum_{c=1}^K |\mathcal{G}_c|$ is the number of genes in clusters, $N_J = \sum_{c=1}^K |\{c: J_c > 0\}|$ is the number of clusters associated with modules and $N_I = \sum_{d=1}^D |\mathcal{A}_d|$ is the total number of QTLs. Finally, the posterior probability of $\{I, J, C\}$ can be written as

$$\begin{aligned} \Pr(I, J, C, \{\mathbf{t}_d\}_{1 \leq d \leq D} | \mathbf{x}, \mathbf{y}) &\propto \prod_{d=1}^D \left[\Pr(\mathbf{x}_{\mathcal{A}_d} | \mathbf{t}_d) \Pr(\mathbf{x}_{\mathcal{A}_d^c} | \mathbf{x}_{\mathcal{A}_d}) \prod_{c: J_c = d} \Pr(\mathbf{y}_{\mathcal{G}_c} | \mathbf{t}_d) \Pr(\mathbf{t}_d) \right] \\ &\times \prod_{c: J_c = 0} \Pr(\mathbf{y}_{\mathcal{G}_c}) \Pr(\mathbf{y}_{\mathcal{G}_0}) \Pr(I, J, C) \end{aligned} \quad (2)$$

For the remainder of this section, we will focus on each model component in details. In the next section, we will discuss the choice of hyper-parameters and introduce an MCMC algorithm to sample from the posterior distribution in (2). For simplicity of description, we will omit the subscript d when discussing a single eQTL module in the following subsections.

2.2 A hierarchical model of gene expression

In this section, we propose a model of gene expression traits that takes into account the random effects of both gene clusters and individual types. For genes in cluster c , given individual types $\mathbf{t} = \{t_i\}_{i=1}^n$, we assume the following hierarchical model:

$$Y_{i,j} | C_j = c \sim N(\tau_{i,c}, \sigma^2), \tau_{i,c} | T_i = t \sim N(\mu_{t,c}, \sigma^2 / \kappa_1), \text{ and } \mu_{t,c} \sim N(0, \sigma^2 / \kappa_2), \quad (3)$$

where $\tau_{i,c}$ is the mean of all the genes in cluster c for individual i , σ^2 is the within-cluster variance for an individual, and κ_1 and κ_2 are higher level scale parameters. The second level model imposes that the $\tau_{i,c}$ of all the individuals of the same type $T = t$ follow another Gaussian distribution with mean $\mu_{t,c}$. Intuitively, κ_2 measures the similarity of “average” gene expression relative to σ^2 between individual types and κ_1 measures the similarity of “average” gene expression relative to σ^2 between individuals with the same individual type. We further assume that the following prior distribution on variance parameters $\Theta = \{\sigma^2, \kappa_1, \kappa_2\}$:

$$\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2), \kappa_1 \sim \chi^2(\nu_1, \sigma_1^2), \text{ and } \kappa_2 \sim \chi^2(\nu_2, \sigma_2^2),$$

where $\{\nu_k, \sigma_k^2\}_{k=0,1,2}$ are hyper-parameters. After integrating out intermediate parameters, we can derive the conditional distribution of $\{Y_{i,j} = y_{i,j}\}_{C_j=c, 1 \leq i \leq n}$ given an individual type partition $\mathbf{t}$ and variance parameters Θ :

$$\Pr(\mathbf{y}_{\mathcal{G}_c} | \mathbf{t}, \Theta) = (2\pi\sigma^2)^{-\frac{nN_c}{2}} Z_{c,\kappa_1,\kappa_2} \exp\left(-\frac{S_{c,\kappa_1,\kappa_2}^2}{2\sigma^2}\right), \quad (4)$$

with

$$Z_{c,\kappa_1,\kappa_2} = \left(\sqrt{\frac{\kappa_1}{N_c + \kappa_1}}\right)^n \left(\prod_{t=1}^{|T|} \sqrt{\frac{(N_c + \kappa_1)\kappa_2}{(N_c + \kappa_1)\kappa_2 + N_c n_t \kappa_1}}\right),$$

and

$$S_{c,\kappa_1,\kappa_2}^2 = \sum_{i=1}^n \left[\sum_{C_j=c} y_{i,j}^2 - \frac{(\sum_{C_j=c} y_{i,j})^2}{N_c + \kappa_1} \right] - \sum_{t=1}^{|T|} \frac{\kappa_1^2 (\sum_{T_i=t} \sum_{C_j=c} y_{i,j})^2}{(N_c + \kappa_1)[(N_c + \kappa_1)\kappa_2 + N_c n_t \kappa_1]}, \quad (5)$$

where $N_c = |\mathcal{G}_c|$ is the number of genes in cluster c , n_t is the number of individuals with individual type $T_i = t$ and $|T|$ is the number of distinct individual types in $\mathbf{t} = \{t_i\}_{1 \leq i \leq n}$. Note that the variance parameters Θ are shared across all gene clusters linked to modules, that is, $\{c : J_c > 0\}$. Instead of analytically marginalizing out variance parameters Θ to obtain $\Pr(\mathbf{y}_{\mathcal{G}_c} | \mathbf{t})$, we augment model (2) with Θ and sample from the joint posterior distribution using a data augmentation procedure described in the Supplementary Materials.

For a gene cluster c not linked to any module, that is, $J_c = 0$, we assume that it follows a hierarchical model with all individuals having the same individual type. Specifically, by assuming $\kappa_1 = 1$, $\kappa_2 = \infty$ and integrating out σ^2 in (4), we have

$$\Pr(\mathbf{y}_{\mathcal{G}_c}) = Z_{c,1,\infty} \frac{\Gamma\left(\frac{nN_c + \nu_0}{2}\right) (\nu_0 \sigma_0^2)^{\frac{\nu_0}{2}}}{[\Gamma(1/2)]^{nN_c} \Gamma(\nu_0/2) \left(S_{c,1,\infty}^2 + \nu_0 \sigma_0^2\right)^{\frac{nN_c + \nu_0}{2}}}. \quad (6)$$

For genes not belonging to any cluster, that is, $\mathcal{G}_0 = \{j : C_j = 0\}$, we assume that their standardized expression levels follow independent standard Gaussian distributions.

2.3 A Dirichlet-multinomial model of QTLs

For a given module, the association indicator $I_k = 1$ if SNP indexed by k is a quantitative trait locus (QTL) linked to given individual type labels $\mathbf{t} = \{t_i\}_{i=1}^n$, and $I_k = 0$ otherwise. We write $\mathcal{A} = \{k : I_k = 1\}$ and let $|\mathcal{A}|$ denote the number of SNPs in $\mathcal{A}$. Conditional on the individual type label $\mathbf{t}$, the distribution of SNPs in $\mathcal{A}$, denoted as $X_{\mathcal{A}}$, is assumed to be

$$X_{\mathcal{A}}|T=t \sim \text{Multinomial} \left(1, \theta_{\mathcal{A}}^{(t)} \right),$$

where $\theta_{\mathcal{A}}^{(t)}$ is a vector with $V^{|\mathcal{A}|}$ elements and each element corresponds to the frequency of observing a particular combination of SNP genotypes from $\mathcal{A}$. We further assume that $\theta_{\mathcal{A}}^{(t)}$ follows the following Dirichlet distribution *a priori*:

$$\theta_{\mathcal{A}}^{(t)} \sim \text{Dirichlet} \left(\frac{\alpha}{V^{|\mathcal{A}|}}, \dots, \frac{\alpha}{V^{|\mathcal{A}|}} \right),$$

where α is a hyper-parameter to be specified. After integrating out $\theta_{\mathcal{A}}^{(t)}$, we can directly write down the probability of observing $\{X_{i,j} = x_{i,j}\}_{1 \leq i \leq n, j \in \mathcal{A}}$ given their individual types $\{T_i = t_i\}_{1 \leq i \leq n}$

$$\Pr(\mathbf{x}_{\mathcal{A}}|\mathbf{t}) = \prod_{t=1}^{|T|} \frac{\Gamma(\alpha)}{\Gamma(\alpha + n_t)} \prod_{h=1}^{V^{|\mathcal{A}|}} \frac{\Gamma(n_t^{(h)} + \frac{\alpha}{V^{|\mathcal{A}|}})}{\Gamma(\frac{\alpha}{V^{|\mathcal{A}|}})}, \quad (7)$$

where n_t is the number of observations with individual type t , $n_t^{(h)}$ is the number of observations with genotype combination h and individual type t and $|T|$ is the number of distinct individual types in $\mathbf{t} = \{t_i\}_{1 \leq i \leq n}$.

The saturated Dirichlet-multinomial model in (7) has an exponentially growing complexity as the number of QTLs increases. We can further enhance our ability in detecting SNPs with weak effects by grouping QTLs into approximately conditionally independent cliques. Specifically, we divide associated SNPs in $\mathcal{A}$ into M groups (M is random), denoted as $\mathcal{A}^{(1)}, \dots, \mathcal{A}^{(M)}$, such that $X_{\mathcal{A}^{(1)}}, \dots, X_{\mathcal{A}^{(M)}}$ are independent conditional on $\mathbf{t}$, that is,

$$\Pr(\mathbf{x}_{\mathcal{A}}|\mathbf{t}) = \prod_{m=1}^M \Pr(\mathbf{x}_{\mathcal{A}^{(m)}}|\mathbf{t}), \quad (8)$$

where each $\Pr(\mathbf{x}_{\mathcal{A}^{(m)}}|\mathbf{t})$ ($m = 1, \dots, M$) is described by a saturated Dirichlet-multinomial distribution in (7). We expand the support of the SNP association indicator I_k from $\{0, 1\}$ to $\{0, 1, 2, \dots\}$, such that $I_k = m$ if $k \in \mathcal{A}^{(m)}$ for $m = 1, 2, \dots$ and $I_k = 0$ if the SNP indexed by k is not associated with the trait. We further assume that the nonzero I_k 's follow a Chinese restaurant process. That is, I_k joins one of non-zero group in $I_{[-k]} = \{I_k : k' \prec k\}$ with probability proportional to the size of that group, and becomes a new group with probability proportional to a pre-specified concentration parameter ω_1 .

Here, we assume that SNPs within the same group interact fully with each other and SNPs in different groups are conditionally independent given individual types. Zhang (2012)

proposed to model the interactions between SNPs using Bayes networks, which can be adopted to further refine the current model.

2.4 Model of background SNPs conditioning on QTLs

To model “background” SNPs in a given module, we consider a Dirichlet-multinomial distribution similar to (7) but without conditioning on individual type T . Given QTLs linked to the module, $X_{\mathcal{A}}$, we use $X_{\mathcal{A}^c}$ to denote the set of background SNPs. We assume that the conditional distribution of $X_{\mathcal{A}^c}$ given $X_{\mathcal{A}}$ is

$$X_{\mathcal{A}^c} | X_{\mathcal{A}} = h \sim \text{Multinomial} \left(1, \theta_{\mathcal{A}^c}^{(h)} \right),$$

where $\theta_{\mathcal{A}^c}^{(h)}$ is a frequency vector with $V|\mathcal{A}^c|$ elements given that QTLs $X_{\mathcal{A}}$ has a particular genotype combination h . We further assume that $\theta_{\mathcal{A}^c}^{(h)}$ follows a Dirichlet prior

$$\theta_{\mathcal{A}^c}^{(h)} \sim \text{Dirichlet} \left(\frac{\alpha_0}{V^p}, \dots, \frac{\alpha_0}{V^p} \right),$$

where α_0 is a hyper-parameter. After integrating out $\theta_{\mathcal{A}^c}^{(h)}$, one can show that the conditional distribution of all SNPs $\mathbf{x}$ given $\mathbf{x}_{\mathcal{A}}$ is given by

$$\Pr(\mathbf{x}_{\mathcal{A}^c} | \mathbf{x}_{\mathcal{A}}) = \frac{\Pr_{\text{null}}(\mathbf{x})}{\Pr_{\text{null}}(\mathbf{x}_{\mathcal{A}})}, \quad (9)$$

with $\Pr_{\text{null}}(\mathbf{x})$ and $\Pr_{\text{null}}(\mathbf{x}_{\mathcal{A}})$ defined as

$$\Pr_{\text{null}}(\mathbf{x}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + n)} \prod_{h'=1}^{V^p} \frac{\Gamma(n^{(h')} + \frac{\alpha_0}{V^p})}{\Gamma(\frac{\alpha_0}{V^p})}, \quad (10)$$

and

$$\Pr_{\text{null}}(\mathbf{x}_{\mathcal{A}}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + n)} \prod_{h=1}^{V|\mathcal{A}|} \frac{\Gamma(n^{(h)} + \frac{\alpha_0}{V|\mathcal{A}|})}{\Gamma(\frac{\alpha_0}{V|\mathcal{A}|})}, \quad (11)$$

where $\mathbf{x} = \mathbf{x}_{\mathcal{A}} \cup \mathbf{x}_{\mathcal{A}^c}$, $n^{(h')}$ is the number of observations with genotype combination h' from SNPs in $\{1, \dots, p\}$ and $n^{(h)}$ is the number of observations with genotype combination h from SNPs in $\mathcal{A}$. Note that (10) and (11) are in the form of Dirichlet-multinomial distribution, and we use the subscript $\Pr_{\text{null}}(\cdot)$ to distinguish the probability under the null model from the probability model for QTLs linked to individual types.

Since our goal is to infer the QTL set $\mathcal{A} = \{k : I_k = 1\}$, we can avoid computing $\Pr_{\text{null}}(\mathbf{x})$ in (10) (which can be computationally intensive when p is large). Specifically, the posterior probability of $I = \{I_k\}_{k=1}^p$ can be written as

$$\Pr(I|\mathbf{t}, \mathbf{x}) \propto \Pr(\mathbf{x}_{\mathcal{A}}|\mathbf{t}) \Pr(\mathbf{x}_{\mathcal{A}^c}|\mathbf{x}_{\mathcal{A}}) \Pr(I) \propto \frac{\Pr(\mathbf{x}_{\mathcal{A}}|\mathbf{t})}{\Pr_{\text{null}}(\mathbf{x}_{\mathcal{A}})} \left(\frac{\eta_I}{1 - \eta_I} \right)^{|\mathcal{A}|}, \quad (12)$$

where $\Pr_{\text{null}}(\mathbf{x})$ is omitted after the “ $\propto$ ” sign since it does not depend on I .

2.5 Block model of linkage disequilibrium

Because of linkage disequilibrium, adjacent SNPs on a chromosome can be highly correlated with a block-wise dependence structure (known as LD blocks). By working with SNP blocks instead of individual SNPs, we can reduce false positives and significantly improve computational efficiencies without sacrificing much statistical power.

Without loss of generality, we assume that SNPs are on the same chromosome and have been sorted according to their locations I_k , that is, $l'_k < l'_k$ for $k' < k$. Suppose the whole genome is partitioned into $|B|$ blocks, denoted as $B = \{\mathcal{L}_b\}_{b=1}^{|B|}$, and let $\mathcal{L}_b$ represent consecutive SNPs in a block. Given a block partition B , we assume that the SNPs in the block $\mathcal{L}_b$ have the distribution:

$$X_{\mathcal{L}_b} \sim \text{Multinomial}(1, \theta_{\mathcal{L}_b}),$$

and

$$\theta_{\mathcal{L}_b} \sim \text{Dirichlet}\left(\frac{\alpha_0}{V|\mathcal{L}_b|}, \dots, \frac{\alpha_0}{V|\mathcal{L}_b|}\right).$$

Then, we can obtain an explicit formula for $\Pr_{\text{block}}(X_{\mathcal{L}_b})$ similar to (11),

$$\Pr_{\text{block}}(\mathbf{x}_{\mathcal{L}_b}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + n)} \prod_{t=1}^{V|\mathcal{L}_b|} \frac{\Gamma\left(n_h + \frac{\alpha_0}{V|\mathcal{L}_b|}\right)}{\Gamma\left(\frac{\alpha_0}{V|\mathcal{L}_b|}\right)},$$

where n_h is the number of observations with genotype combination h from SNPs in $\mathcal{L}_b$. Here, we use $\Pr_{\text{block}}(\cdot)$ to denote the probability of observing SNPs $\mathbf{x}_{\mathcal{L}_b}$ in block h . To reduce model complexity, we approximate the distribution of background SNPs using a block-based model. Specifically, given the block partition B , the SNPs in different blocks are assumed to be independent, that is,

$$\Pr_{\text{block}}(\mathbf{x}|B) = \prod_{j=1}^{|B|} \Pr_{\text{block}}(\mathbf{x}_{\mathcal{L}_b}).$$

We assign a prior probability $\Pr(B)$ by assuming that there is a probability of π_b to start a block at a genomic locus *a priori*. Then we can use a dynamic programming algorithm to calculate the maximum a posteriori (MAP) estimates of the block structure (see Supplementary Materials for details).

Given LD blocks $B = \{\mathcal{L}_b\}_{b=1}^{|B|}$, we impose an additional restriction on SNP association indicators $\{I_k\}_{k=1}^p$ such that $\sum_{k \in \mathcal{L}_b} I_k \leq 1$, that is, at most one SNP in a block can be associated with the given module.

3 MCMC sampling algorithm and implementation

3.1 Choice of hyper-parameters

There are several hyper-parameters that need to be specified, including the number of gene clusters K , the prior probabilities $\{\eta_C, \eta_I, \eta_D\}$, hyper-parameters $\{\nu_j, \sigma_j^2\}_{j=0}^2$ for variances in the hierarchical model, concentration parameters $\{\omega_0, \omega_1\}$ in the Chinese restaurant process, α_0 in the Dirichlet-multinomial model and π_b on the number of LD blocks.

In practice, we recommend choosing the number of gene clusters K to be moderately large (say 100 to 500) so that we can capture the detailed correlation structure among gene expressions. Priors η_I and π_b should be chosen based on prior knowledge. In the yeast data set, we assume there are 5 SNPs associated with each module *a priori*, and set $\eta_I = 5/p$ and $\pi_b = 100/p$ corresponding to about 100 blocks. Furthermore, we use $\alpha_0 = 1$, the Jeffreys' prior when there are two types of SNPs on each locus. Finally, we find that our results are not sensitive to the choice of other hyper-parameters and set $\eta_C = \eta_I = 0.05$, $\nu_j = \sigma_j^2 = 1$ ($j = 0, 1, 2$) and $\omega_0 = \omega_1 = 1$ for the Chinese restaurant process priors on individual types and QTL groups.

A SNP k ($k = 1, \dots, p$) is declared to be associated with a module d ($d = 1, \dots, D$) if its corresponding marginal posterior probability of association, *i.e.* $\Pr(I_{k,d} = 1 | \mathbf{x}, \mathbf{y})$, is greater than a given threshold, which is chosen as 0.5 in this paper. One may also choose a desired threshold to control false discovery rate under the Bayesian paradigm such as the direct posterior probability approach in Newton *et al.* (2004).

3.2 Preprocessing and initialization

There are several data processing steps before applying the BP2 model. First, if there are unobserved SNP genotypes in a data set, one can use existing tools such as IMPUTE2 (Howie *et al.*, 2009) or MaCH (Li *et al.*, 2010) to impute the missing values. We suggest filtering out SNPs with small minor allele frequencies (say below 5%) in the data set. Second, we remove genes with small expression variations among individuals (e.g. genes whose expression variance is smaller than 10% of median variance of all genes) before applying quantile normalization on gene expression. Then, we standardize the expression level of each gene to have zero mean and unit variance.

Given pre-processed SNP and gene expression data as inputs, BP2 model starts by initializing LD blocks, gene clusters and their module memberships according to the following procedures:

1. According to the block model introduced in Section 2.5, we use the dynamic programming algorithm described in the Supplementary Materials to partition the whole genome into blocks of highly correlated SNPs.
2. Initialize K gene clusters based on model (6) in Section 2.2 with all individuals having the same individual type. Note that the hierarchical model can only group positively correlated genes into the same cluster.
3. Within each initialized gene cluster, rank individuals by their average expression levels. We further group gene clusters with correlated ranks into a “super-cluster”. Specifically, define a super-cluster $\mathcal{C}$ as a collection of gene clusters and a similarity measure between two super-clusters $\mathcal{C}_1$ and $\mathcal{C}_2$ as

$$\rho(\mathcal{C}_1, \mathcal{C}_2) = \max_{c_1 \in \mathcal{C}_1, c_2 \in \mathcal{C}_2} |r_s(c_1, c_2)|,$$

where $r_s(c_1, c_2)$ is the Spearman’s rank correlation between the ranks of average expression levels in two clusters c_1 and c_2 . Given a pre-specified threshold ρ_0 (e.g. $\rho_0 = 0.6$), we determine super-clusters as follows: (1) start with K initial super-clusters and each of them contains a single gene cluster; (2) iteratively select two most similar superclusters with similarity measure $\rho_{\max}$ and merge them into one; (3) terminate when $\rho_{\max} < \rho_0$ and output the final list of super-clusters.

4. We choose the number of modules D to be the number of super-clusters determined in the previous step, and link all gene clusters in the d th super-cluster ($d = 1, \dots, D$) to a module d by letting $J_c = d$.

3.3 MCMC sampling algorithm

After initialization, we iteratively update parameters of interest according to their posterior distributions in (2) through the following steps:

Algorithm 1

-
- *Step 1:* Sample gene cluster indicators for each gene, $\{C_j\}_{1 \leq j \leq q}$. For genes $j = 1, 2, \dots, q$, iteratively update C_j conditioning on $C_{[-j]} = \{C_{j'}: j' \neq j\}$, individual type partitions $\{T_d\}_{1 \leq d \leq D}$ and variance parameters Θ .
 - *Step 2:* Sample module memberships of gene clusters, $\{J_c\}_{1 \leq c \leq K}$. For gene clusters $c = 1, 2, \dots, K$, iteratively update J_c conditioning on $J_{[-c]}$, $\{T_d\}_{1 \leq d \leq D}$ and variance parameters Θ .
 - *Step 3:* For module $d = 1, 2, \dots, D$, sample SNP association indicators in each module d , i.e. $\{I_{k,d}\}_{1 \leq k \leq p}$. For SNP blocks $b = 1, \dots, |B|$, either choose the SNP $k \in \mathcal{L}_b$ with $I_{k,d} > 0$ or randomly select a SNP $k \in \mathcal{L}_b$ from the block if $I_{k,d} = 0$ for all $k \in \mathcal{L}_b$. Conditioning on $\{I_{k',d}: k' \neq k\}$ and $\{T_d\}_{1 \leq d \leq D}$, update $I_{k,d}$ according to a Metropolis-Hasting algorithm with acceptance ratio proportional to its posterior probability and the size of the block.
 - *Step 4:* Conditioning on $\{I, J, C\}$, sample the variance parameters $\Theta = \{\sigma^2, \kappa_1, \kappa_2\}$ according to the data augmentation procedure described in the Supplementary Materials.

- *Step 5:* For module $d = 1, \dots, D$, sample individual types $\mathbf{t}_d$. For individuals $i = 1, \dots, n$, iteratively update $T_{i,d}$ conditioning $\{I_{i',d}: i' \neq i\}$ indicators $\{I, J, C\}$ and variance parameters Θ .

On a typical yeast data set with ~ 100 individuals, ~ 3000 SNPs and ~ 4000 genes, the above MCMC algorithm takes about an hour to finish 500 iterations on a PC. When applying the method to extremely large data sets, one can potentially speed up the computation by parallel updating each module independently after initialization. Diagnostics of MCMC convergence in simulation studies are presented in the Supplementary Materials.

4 Simulation studies

In this section, we compare the performance of the Bayesian hierarchical partition model, BP2, with the original BP method in Zhang *et al.* (2010) and other eQTL methods. The first simulation study is designed the same way as in Zhang *et al.* (2010), where genes in the same module are positively correlated. To mimic more complex gene expression patterns in real data, in the second simulation study, we modify the original design to allow genes in the same module to be either positively or negatively correlated.

We analyze the simulated data sets using five methods: (1) the original BP method proposed by Zhang *et al.* (2010), referred to as BP1; (2) the new method developed in this paper, referred to as BP2; (3) a two-stage stepwise regression method applied to individual gene expression proposed by Storey *et al.* (2005), referred to as SR; (4) iBMQ (Scott-Boyer *et al.*, 2012; Imholte *et al.*, 2013), an integrated hierarchical Bayesian regression model that jointly models expression levels of all genes conditioning on all SNPs to detect eQTLs; (5) a two-stage stepwise regression method applied to the first principle component (PC) of expression levels of known genes in each module, referred to as PCA. The SR method has two stages: in the first stage, it identifies the most significant marker for each gene expression trait based on the one-gene-one-marker regression model. It then proceeds to find the next most significant marker conditional on the previous detected marker for each gene. Permutation tests over all genes are carried out in each stage to control the overall false discovery rate (FDR). The iBMQ method is based on a Bayesian sparse regression model of gene expression given SNPs. Instead of explicitly modeling gene expression correlations, it assumes that gene expression levels are conditionally independent given SNP-gene association indicators, and borrows information across all genes by assuming a common prior on association probabilities of each SNP. The PCA method assumes that the *true* genes in each module are known, and serves as an oracle benchmark for the SR method.

4.1 Simulation with positively correlated genes

As with Zhang *et al.* (2010), we simulated 120 individuals with 500 binary markers and 1000 expression traits in the context of inbred cross of haploid strains. Given the haploid nature of the segregants, 500 binary markers are equally spaced on 20 chromosomes, each of length 100cM, using the qtl package in R. There are 8 modules (denoted as A, B, ..., H), each consisting of 40 genes and 2 associated markers, simulated from different epistasis models based on the linear regression framework. The associated markers in each module

are randomly selected and do not overlap. Note that the generative models in our simulation studies are different from the posited Bayesian partition model.

To mimic inter-module correlations of the genes in real gene expression data, we first generated a core gene in each module according to the corresponding models depicted in Table 1. In each model, $\varepsilon \sim N(0, \sigma_e^2)$ represents the environmental noise. The regression coefficient β in each model was chosen such that the percentage of total variance explained by all the relevant SNPs is 60% for the core gene. After generating the core gene, we simulated the gene expression traits in each module independently from a Gaussian model conditional on the core gene so that they have a given average correlation to the core gene. In this simulation study, we fixed the average correlation for genes within each module with the core gene at 0.5 across all eight modules. Finally, we calculated the percentage of variation explained by the true model averaged over all genes in a module as listed in the third column of Table 1. For example, for each gene in module B we calculated the sum of squares of the gene expression for all 120 samples (SS_{total}) and the residual sum of squares (SS_{res}) within the two sample groups: those with $x_1 = x_2$ and those with $x_1 \neq x_2$. As a result,

the percentage of variation explained by the true model for this gene is $1 - \frac{SS_{\text{res}}}{SS_{\text{total}}}$.

To get a better understanding of the signal strength in each module, we divided the total genetic variance for a two-locus model into three components: the genetic variance at locus 1, the genetic variance at locus 2, and the epistatic (interaction) variance using the classical analysis of variance (Fisher, 1919; Cockerham, 1954; Tiwari and Elston, 1997). The relative percentages of three variance components are listed as the last three columns in Table 1, which add up to one. The details of ANOVA decompositions is given in the Supplementary Materials.

We apply four methods, BP1, BP2, SR, iBMQ and PCA, to 100 simulated data sets. To run BP1, we need to specify the number of modules and we give BP1 some advantage by using the true number, $D = 8$. For BP2, we assume that we do not know the true number of gene clusters or modules and use a larger number of gene clusters, $K = 20$. The number of modules is determined by the procedure described in Section 3.1. Under a range of thresholds on absolute Spearman's correlations $\rho_0 \in [0.5, 0.8]$, we were able to correctly determine the number of modules in most of the simulations. We choose $\rho_0 = 0.6$ to obtain the following results. Diagnostics of MCMC convergence on simulated data sets are presented in the Supplementary Materials.

The receiver operating characteristic (ROC) curves in Figure 2 compare true positives, the total number of the true gene-marker pairs detected, and false positives, the total number of unrelated gene-marker pairs falsely selected, at varying thresholds. Figure 3 further compares true positives and false positives of different methods in each module. As shown from the ROC curves in Figure 2 and Figure 3, in modules that have strong marginal but weak interactive effects, BP2 performed almost as good as the PCA method based on the stepwise regression, even though the latter has already been given the true set of genes in each module to start with. In modules that have weak marginal but strong interactive effects (module B, D and H), BP2 was more powerful than the PCA method in detecting epistasis

effects. When the true genes in modules are not given, the stepwise method SR based on the one-gene-one-marker regression model had the lowest detection rate, especially when there are strong epistasis effects. Moreover, BP2 achieved consistently and significantly higher power in detecting eQTLs (gene-marker pairs) compared to the iBMQ method and the original model, BP1. There are several reasons for the excellent performances of BP2. First, BP2 uses a more efficient algorithm to partition individuals, and a more flexible model of the dependence structure between gene expression and SNPs. Second, we aggregate information from all co-regulated genes in a module and improve the signal strength of eQTLs. Third, by using a joint model of interactive markers and an iteratively sampling approach, we significantly increase the power in detecting markers with weak marginal but strong interactive effects compared to the stepwise methods that select one marker at a time.

4.2 Simulation with mixed correlations

Our second simulation studies the performance of different methods when there are both positively and negatively correlated genes in the same module. The data generation process is the same as in the previous simulation except that a random sign is multiplied to the simulated expression of each gene. Since the original BP model cannot capture negatively correlated genes in the same module, we use 16 (the number of gene groups with positively correlated gene expression) instead of 8 as the “true” number of modules for BP1. For BP2, we again specify the number of gene clusters as 20 and initialize the modules using the procedure described in Section 3.1 with threshold $\rho_0 = 0.6$. The aggregated ROC curves of different methods are shown in Figure 4 and the ROC curves in each module are shown in Figure 5.

As expected, the original Bayesian partition model, BP1, has a lower power in the second simulation compared to its performance in the first simulation. Although we increased the number of modules in BP1 from 8 to 16 in order to capture all relevant genes, the separation of negatively correlated genes into different modules (a module only contains positively correlated genes in BP1) weakened the signal strength of gene expression in determining individual type partitions. The lower detection rate of BP1 is more evident in Module B, D and H, when an informative partition of individuals becomes critical in detecting SNPs with weak marginal but strong interactive effects. In contrast, BP2 is able to combine negatively correlated genes in the same module and shows consistently excellent performances in both simulations. In modules E, F, and G where the major marker explains more than 70% of the genetic variation, the PCA method, which starts with the true gene-module assignments and uses stepwise regression to detect markers, outperformed BP2. In Module A and C where the marginal effects of the two marker are almost the same, BP2 and the PCA method have comparable performances. In Module B, D and H where no or very weak marginal effect is present and genetic variations are mainly explained by the epistasis, BP2 achieved significantly better power than the PCA method, even though the latter has a full knowledge of genes in each module. The results of SR and iBMQ are similar to those in the previous section. This is because iBMQ assumes that gene expression levels are conditionally independent given SNP-gene association indicators and its performance is not affected by the multiplication of a random sign.

5 Yeast data analysis

In this section, we present an application of the BP2 model to a yeast data set with $p = 2957$ markers and $q = 3662$ gene expression profiles from $n = 112$ yeast (*S. cerevisiae*) segregants (Brem and Kruglyak, 2005; Zhang *et al.*, 2010). We set the number of gene clusters $K = 200$ and the number of modules $D = 100$ in this study. Because markers in the yeast data set are very densely distributed, adjacent markers are highly correlated. After MCMC sampling, markers adjacent to the truly linked marker often dilute the posterior probability for the true marker-module linkage. To counter this problem, we first specify a window centered at each marker so that markers inside the window are in high LD with the marker in the center. The posterior probabilities of all markers in the window are summed up and regarded as the modified posterior probability of the central marker. The markers with peak probabilities exceeding the given threshold are selected and all other markers in the corresponding windows are masked out. We choose the window size to contain 5 markers and 0.5 as the threshold for modified posterior probabilities to determine the module membership of a marker. Among 100 modules, 36 modules are not associated with any marker above the threshold, 52 modules are associated with a single markers, 11 modules are associated with two markers and 1 module is associated with three markers.

Figure 6 shows an example of a module linked to a single marker on Chromosome XII. The genes in the module are grouped into two positively correlated gene clusters with negative correlations between two clusters. The functional annotation of each gene cluster is shown on top of the figure. Out of the 14 genes in the module, nine of them are physically located adjacent to the SNP and are cis-acting eQTLs. The other five genes are located on different chromosomes and are trans-acting eQTLs.

Figure 7 shows a module linked to two SNPs. There are two gene clusters in the module with a total of 27 genes, most of which are related to the sexual reproduction process in yeast. Nine out of 27 genes are located near the marker YCR041W on Chromosome III. The other 18 genes are not located in adjacent to either marker. Box-plots of average gene expression in two gene clusters under different genotype combinations are shown in Figure 8. From Figure 7 and Figure 8, we can see that the marker YCR041W has a primary regulatory effect and divides individuals into two separate groups in both gene clusters. The secondary marker YHL007C further divides the low-expression individuals into two subgroups.

Figure 9 shows another example of a module that is linked to two SNPs. The three gene clusters in the module exhibit more complicated gene expression patterns and all of them are involved in organic acid biosynthetic process. Both SNPs are trans-eQTLs. Individuals with genotype combination (1, 0) from two markers have low expression in the first gene cluster and high expression in the second gene cluster, and individuals with genotype combination (1, 1) have relatively high expression in the third gene cluster.

In the example shown in Figure 10, we identified a module linked to three SNPs. The module consists of four genes with functions related to oxidation-reduction and dehydrogenase. The three SNPs in the module are trans-eQTLs co-localized with other

genes involved in oxidation-reduction, dehydrogenase and ATP-binding respectively. From the heatmap in Figure 10, we can see that when the three SNPs have genotype combination (1, 1, 1), the four trans-acting genes in the module will have relatively higher expression compared with individuals with other genotype combinations.

6 Discussion

We have described a full Bayesian model for identifying pleiotropic and epistasis effects in eQTL studies. Novelities of the Bayesian hierarchical partition model, BP2, are threefold. First, it improves signal strength by aggregating information from correlated gene clusters and allowing negatively correlated genes to be included in the same module. Second, it directly accounts for dependence structures of SNPs by modeling them as linkage disequilibrium (LD) blocks. Third, by integrating out intermediate parameters in the hierarchical model of gene clusters and modeling variance/scale parameters as random effects, BP2 allows for adaptive estimation of gene clusters and more efficient computations. Simulation studies have demonstrated that BP2 achieved a significantly improved power in detecting eQTLs compared to the original BP1 method and regression-based methods including two-stage stepwise regression and hierarchical Bayesian regression. We applied BP2 to analyzing a yeast eQTL dataset and found numerous interesting pleiotropic and epistatic modules. A particular strength of BP2 our method is its ability to detect epistatic effects with high power when the marginal effects are weak, addressing a key weakness of other eQTL mapping methods. The software that implements the proposed method can be downloaded from <http://www.people.fas.harvard.edu/~junliu/BP/>.

Further improvements of the model are possible. First, Zhang (2012) proposed a refined model of the interactions between SNPs using Bayes networks, which can be incorporated into our Bayesian partition model. Second, although the current BP2 model assumes that the missing SNP genotypes have been imputed in a previous step, the Bayesian framework can be extended to directly model missing data. Third, human SNP data often involve 0.5 million to 2.5 million of SNPs, parallelizations, *e.g.*, updating each module independently after initialization, can greatly speed up the computations of BP2 on such high-dimensional data sets. Last but not least, using gene expression data from multiple tissues, the BP2 model can be further generalized to study tissue common and tissue specific eQTLs. We are currently collaborating with scientists in the Genotype-Tissue Expression (GTEx) project, which aims to comprehensively survey genetic regulation of gene expression in multiple human tissues.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Jun S. Liu was supported in part by NSF grants DMS-1120368 and DMS-1007762, and by Shenzhen Special Fund for Strategic Emerging Industry (No.ZD201111080127A).

References

- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1995;89–300.
- Bing N, Hoeschele I. Genetical genomics analysis of a yeast segregant population for transcription network inference. *Genetics*. 2005; 170(2):533–542. [PubMed: 15781693]
- Brem RB, Kruglyak L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences of the United States of America*. 2005; 102(5):1572–1577. [PubMed: 15659551]
- Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. *Science*. 2002; 296(5568):752–755. [PubMed: 11923494]
- Bystrykh L, Weersing E, Dontje B, Sutton S, Pletcher MT, Wiltshire T, Su AI, Vellenga E, Wang J, Manly KF, et al. Uncovering regulatory pathways that affect hematopoietic stem cell function using ‘genetical genomics’. *Nature Genetics*. 2005; 37(3):225–232. [PubMed: 15711547]
- Chen Y, Zhu J, Lum PY, Yang X, Pinto S, MacNeil DJ, Zhang C, Lamb J, Edwards S, Sieberts SK, et al. Variations in dna elucidate molecular networks that cause disease. *Nature*. 2008; 452(7186):429–435. [PubMed: 18344982]
- Chesler EJ, Lu L, Shou S, Qu Y, Gu J, Wang J, Hsu HC, Mountz JD, Baldwin NE, Langston MA, et al. Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nature Genetics*. 2005; 37(3):233–242. [PubMed: 15711545]
- Chun H, Kele S. Expression quantitative trait loci mapping with multivariate sparse partial least squares regression. *Genetics*. 2009; 182(1):79–90. [PubMed: 19270271]
- Cockerham CC. An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics*. 1954; 39(6):859. [PubMed: 17247525]
- Fisher RA. The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*. 1919; 52(02):399–433.
- Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*. 2009; 5(6):e1000529. [PubMed: 19543373]
- Hubner N, Wallace CA, Zimdahl H, Petretto E, Schulz H, Maciver F, Mueller M, Hummel O, Monti J, Zidek V, et al. Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nature Genetics*. 2005; 37(3):243–253. [PubMed: 15711544]
- Imholte GC, Scott-Boyer M-P, Labbe A, Deschepper CF, Gottardo R. ibmq: a r/bioconductor package for integrated bayesian modeling of eqtl data. *Bioinformatics*. 2013; 29(21):2797–2798. [PubMed: 23958729]
- Jiang C, Zeng Z-B. Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics*. 1995; 140(3):1111–1127. [PubMed: 7672582]
- Kendziorski C, Chen M, Yuan M, Lan H, Attie A. Statistical methods for expression quantitative trait loci (eqtl) mapping. *Biometrics*. 2006; 62(1):19–27. [PubMed: 16542225]
- Lan H, Chen M, Flowers JB, Yandell BS, Stapleton DS, Mata CM, Mui ET-K, Flowers MT, Schueler KL, Manly KF, et al. Combined expression trait correlations and expression quantitative trait locus mapping. *PLoS Genetics*. 2006; 2(1):e6. [PubMed: 16424919]
- Lander ES, Botstein D. Mapping mendelian factors underlying quantitative traits using rflp linkage maps. *Genetics*. 1989; 121(1):185–199. [PubMed: 2563713]
- Li H, Lu L, Manly KF, Chesler EJ, Bao L, Wang J, Zhou M, Williams RW, Cui Y. Inferring gene transcriptional modulatory relations: a genetical genomics approach. *Human Molecular Genetics*. 2005; 14(9):1119–1125. [PubMed: 15772094]
- Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*. 2010; 34(8):816–834. [PubMed: 21058334]
- Liu, JS. Monte Carlo Strategies in Scientific Computing. Springer; 2008.
- Mangin B, Thoquet P, Grimsley N. Pleiotropic qtl analysis. *Biometrics*. 1998; 54(1):88–99.

- Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG. Genetic analysis of genome-wide variation in human gene expression. *Nature*. 2004; 430(7001):743–747. [PubMed: 15269782]
- Newton MA, Noueiry A, Sarkar D, Ahlquist P. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*. 2004; 5(2):155–176. [PubMed: 15054023]
- Schadt EE, Monks SA, Drake TA, Lusk AJ, Che N, Colinao V, Ruff TG, Milligan SB, Lamb JR, Cavet G, et al. Genetics of gene expression surveyed in maize, mouse and man. *Nature*. 2003; 422(6929):297–302. [PubMed: 12646919]
- Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, GuhaThakurta D, Sieberts SK, Monks S, Reitman M, Zhang C, et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics*. 2005; 37(7):710–717. [PubMed: 15965475]
- Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, Kasarskis A, Zhang B, Wang S, Suver C, et al. Mapping the genetic architecture of gene expression in human liver. *PLoS Biology*. 2008; 6(5):e107. [PubMed: 18462017]
- Scott-Boyer MP, Imholte GC, Tayeb A, Labbe A, Deschepper CF, Gottardo R. An integrated hierarchical bayesian model for multivariate eqtl mapping. *Statistical applications in genetics and molecular biology*. 2012; 11(4)
- Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*. 2003; 100(16):9440–9445.
- Storey JD, Akey JM, Kruglyak L. Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biology*. 2005; 3(8):e267. [PubMed: 16035920]
- Tiwari HK, Elston RC. Deriving components of genetic variance for multilocus models. *Genetic Epidemiology*. 1997; 14(6):1131–1136. [PubMed: 9433636]
- Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, Mackelprang R, Kruglyak L, et al. Trans-acting regulatory variation in *saccharomyces cerevisiae* and the role of transcription factors. *Nature Genetics*. 2003; 35(1):57–64.
- Zhang W, Zhu J, Schadt EE, Liu JS. A bayesian partition method for detecting pleiotropic and epistatic eqtl modules. *PLoS Computational Biology*. 2010; 6(1):e1000642. [PubMed: 20090830]
- Zhang Y. A novel bayesian graphical model for genome-wide multi-snp association mapping. *Genetic Epidemiology*. 2012; 36(1):36–47. [PubMed: 22127647]
- Zhang Y, Liu JS. Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics*. 2007; 39(9):1167–1173. [PubMed: 17721534]
- Zhu J, Lum P, Lamb J, GuhaThakurta D, Edwards S, Thieringer R, Berger J, Wu M, Thompson J, Sachs A, et al. An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenetic and Genome Research*. 2004; 105(2–4):363–374. [PubMed: 15237224]
- Zhu J, Zhang B, Smith EN, Drees B, Brem RB, Kruglyak L, Bumgarner RE, Schadt EE. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature Genetics*. 2008; 40(7):854–861. [PubMed: 18552845]

Figure 1.
An illustration of an eQTL module with a single gene cluster.

Simulation I

Figure 2.

The aggregated ROC curves that compare true positives, the total number of the true gene-marker pairs detected, and false positives, the total number of unrelated gene-marker pairs falsely selected, of different methods under simulation in Section 4.1. BP1: the original Bayesian partition model (Zhang *et al.*, 2010); BP2: the Bayesian partition model proposed in this paper; SR: a two-stage stepwise method on the one-gene-one-marker regression model (Storey *et al.*, 2005); PCA: a two-stage stepwise method based on the principle component analysis of true genes in each module (oracle benchmark for SR).

Figure 3.

The ROC curves that compare true positives, the total number of the true gene-marker pairs detected, and false positives, the total number of unrelated gene-marker pairs falsely selected, of different methods in each module under simulation in Section 4.1.

Simulation II

Figure 4.

The aggregated ROC curves that compare true positives, the total number of the true gene-marker pairs detected, and false positives, the total number of unrelated gene-marker pairs falsely selected, of different methods under simulation in Section 4.2.

Figure 5.

The ROC curves that compare true positives, the total number of the true gene-marker pairs detected, and false positives, the total number of unrelated gene-marker pairs falsely selected, of different methods in each module under simulation in Section 4.2.

Figure 6. Heatmap for gene expression in a module linked to a single marker (NLR058C) on Chromosome XII. Individuals are divided into two groups according to the genotype (0 or 1) of the SNP. Each column represents the expression level of a gene across individuals. High- and low-expression levels are represented by red and blue, respectively.

Figure 7. Heatmap for gene expression in a module linked to two markers on Chromosome III and VIII. Individuals are divided into four groups according to the genotype combinations of the two markers. Each column represents the expression level of a gene across individuals. High- and low-expression levels are represented by red and blue, respectively.

Figure 8.
Box-plots of average gene expression values under different genotype combinations from two gene clusters in Figure 7.

Figure 9. Heatmap for gene expression in a module linked to two markers. Individuals are divided into four groups according to the genotype combinations of the two markers. Each column represents the expression level of a gene across individuals. High- and low-expression levels are represented by red and blue, respectively.

Figure 10. Heatmap for gene expression in a module linked to three markers. Individuals are divided into eight groups according to the genotype combinations of the two markers. Each column represents the expression level of a gene across individuals. High- and low-expression levels are represented by red and blue, respectively.

Table 1

Simulation design and genetic variance decomposition

Module	Model ¹	% of Var. ²	Locus 1 ³	Locus 2 ⁴	Epistasis ⁵
A	$Y = \beta_1 I_{x_1=1 \text{ or } x_2=1} + \epsilon$	0.166	0.345	0.342	0.313
B	$Y = \beta_1 I_{x_1=x_2} + \epsilon$	0.166	0.058	0.054	0.888
C	$Y = 2\beta_1 I_{x_1=1 \text{ or } x_2=1} + \beta x_1 x_2 + \epsilon$	0.166	0.461	0.445	0.094
D	$Y = \beta_1 I_{x_1=0, x_2=1} + 2\beta_1 I_{x_1=1, x_2=0} + \epsilon$	0.166	0.119	0.116	0.765
E	$Y = \beta x_1 + \beta x_1 x_2 + \epsilon$	0.171	0.749	0.138	0.113
F	$Y = 2\beta x_1 + \beta x_2 + \epsilon$	0.168	0.736	0.216	0.048
G	$Y = 2\beta x_1 + \beta I_{x_1=x_2} + \epsilon$	0.170	0.743	0.058	0.199
H	$Y = 2\beta_1 I_{x_1=0, x_2=1} + 1.5\beta_1 I_{x_1=1, x_2=0} + 0.5\beta_1 I_{x_1=1, x_2=1} + \epsilon$	0.165	0.135	0.053	0.812

¹Regression models that were used to generate the core gene in each module.

²Average percentage of variations of genes in the module explained by the true model.

³Average percentage of genetic variance explained by the first locus.

⁴Average percentage of genetic variance explained by the second locus.

⁵Average percentage of genetic variance explained by epistasis.