A Bayesian Partition Method for Detecting Pleiotropic and Epistatic eQTL Modules

Citation

Published Version
doi:10.1371/journal.pcbi.1000642

Permanent link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:4453998

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Open Access Policy Articles, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#OAP

Share Your Story
The Harvard community has made this article openly available. Please share how this access benefits you. Submit a story.

Accessibility
A Bayesian Partition Method for Detecting Pleiotropic and Epistatic eQTL Modules

Wei Zhang1, Jun Zhu2,3, Eric E. Schadt3,4, Jun S. Liu5

1 UBS Equities, Stamford, Connecticut, United States of America, 2 Rosetta Inpharmatics, LLC, Merck & Co., Inc., Seattle, Washington, United States of America, 3 Sage Bionetworks, Seattle, Washington, United States of America, 4 Pacific Biosciences, Menlo Park, California, United States of America, 5 Department of Statistics, Harvard University, Cambridge, Massachusetts, United States of America

Abstract

Studies of the relationship between DNA variation and gene expression variation, often referred to as “expression quantitative trait loci (eQTL) mapping”, have been conducted in many species and resulted in many significant findings. Because of the large number of genes and genetic markers in such analyses, it is extremely challenging to discover how a small number of eQTLs interact with each other to affect mRNA expression levels for a set of co-regulated genes. We present a Bayesian method to facilitate the task, in which co-expressed genes mapped to a common set of markers are treated as a module characterized by latent indicator variables. A Markov chain Monte Carlo algorithm is designed to search simultaneously for the module genes and their linked markers. We show by simulations that this method is more powerful for detecting true eQTLs and their target genes than traditional QTL mapping methods. We applied the procedure to a data set consisting of gene expression and genotypes for 112 segregants of S. cerevisiae. Our method identified modules containing genes mapped to previously reported eQTL hot spots, and dissected these large eQTL hot spots into several modules corresponding to possibly different biological functions or primary and secondary responses to regulatory perturbations. In addition, we identified nine modules associated with pairs of eQTLs, of which two have been previously reported. We demonstrated that one of the novel modules containing many daughter-cell expressed genes is regulated by AMN1 and BPH1. In conclusion, the Bayesian partition method which simultaneously considers all traits and all markers is more powerful for detecting both pleiotropic and epistatic effects based on both simulated and empirical data.

Introduction

Studies in the genetics of gene expression combine gene expression and genotype data in segregating populations to detect loci linked to variations in RNA levels. These loci are referred to as expression quantitative trait loci (eQTL). To date, eQTL studies have been pursued in a number of species ranging from yeast to mouse and human [1–3]. A common theme of these studies is to treat thousands of gene expression values as quantitative traits and conduct QTL mapping for all of them.

Most eQTL studies are based on linear regression models [4] 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 the association. In the context of multiple traits and markers, procedures such as false discovery rate (FDR) controls [5] can be used to quantify family-wise error rates. Despite the success of this type of regression approach, a number of challenging problems remain. First, these methods can not easily assess the joint effect of multiple markers beyond additive effects. Storey et al. [5] developed a step-wise regression method to find eQTL pairs, then Zhou and Zeng improved it [6]. This procedure, however, tends to miss eQTL pairs with small marginal effects but a strong interaction effect. There are methods for detecting epistatic effects without main marginal effects [7–8]. However, their applications are limited to a few clinical traits instead of thousands of expression traits due to computational constraints. Second, there are often strong correlations among expression levels for certain groups of genes, partially reflecting co-regulation of genes in biological pathways that may respond to common genetic loci and environmental perturbations [2,9–11]. Previous findings of eQTL “hot spots”, i.e., loci affecting a larger number of expression traits than expected by chance, and their biological implications further enhance this notion and highlight the biological importance of finding such gene “modules”. Mapping genetic loci for multiple traits simultaneously is more powerful than mapping single traits at a time [12]. Although for a known small set of correlated traits, one can conduct QTL mapping for the principal components [13], this method 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, and then fit mixture models to clusters of genes [14]. The eQTL mapping then depends on whether the distance metric used by the clustering method is appropriate, whether the method can find the right number of clusters.

We address these issues by modeling the joint distribution of all genes and all markers simultaneously. Under a Bayesian framework, we introduce three sets of latent indicator variables.


Editor: Gary D. Stormo, Washington University School of Medicine, United States of America

Received August 31, 2009; Accepted December 15, 2009; Published January 15, 2010

Copyright: © 2010 Zhang et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by the US National Institutes of Health’s (http://www.nih.gov) grants R01-HG02518-02 and R01-GM078990. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: JZ and EES worked for Merck and own Merck’s stocks.

* E-mail: jliu@stat.harvard.edu
Author Summary

Genome-wide association studies (GWAS) have yielded several causal genes for many human diseases. However, the mechanisms underlying how DNA variations affect disease phenotypes have not been well understood in many cases. Gene expression is intermediate between DNA and clinical endpoints. Linking DNA variation and gene expression variation, often referred to as “expression quantitative trait loci (eQTL) mapping”, has yielded clues of mechanisms and pathways by which DNA variations impact phenotypes. Because of the large number of genes and genetic markers in such analyses, it is extremely challenging to discover how a small number of eQTLs interact with each other to affect mRNA expression levels for a set of co-regulated genes. We present a Bayesian method to identify genetic interactions and more eQTLs by treating co-expressed genes as a module. Our method provides a tool to study genetic interactions in human disease models.

Results

Overview of Bayesian partition method

We define a module as a set of gene expression traits (referred to as “genes” henceforth) and a set of genetic markers (e.g., SNPs) such that the variation of the gene expression traits is associated with the variation of the markers, as shown in Figure 1. This association between multiple genes and markers is characterized by a latent indicator variable, individual type, conditional on which the trait and marker variables are independent of each other. The individual type latent variable can be viewed as representing a certain combination of markers that induces changes in expressions of a certain set of genes across different individual types. In the simplest case with a single marker, the individual type could correspond to a dominant genetic model, as illustrated in Figure 2A. In this instance, our model is mathematically equivalent to the regression model (Figure 2B).

For the case of two markers associated with gene expression traits, there could be two to nine individual types (various genotype combinations). Figure 2C illustrates a case with three individual types: 1) high expression values associated with red-colored genotype combinations, 2) medium expression values with blue-colored combinations, and 3) low expression values with green-colored combinations. The goal of the Bayesian partition method is to simultaneously partition genes and SNPs into modules. The details of the Bayesian partition model are described in the Methods section.

Simulation studies

To test the effectiveness of our method, we simulated 120 individuals with 500 binary markers and 1000 expression traits in the context of inbred cross of haploid strains. There are eight modules (summarized in Table 1), each consisting of 40 genes, simulated from different epistasis models based on the linear regression framework, which is different from the posited Bayesian model in our analysis. The genotypic means and frequencies for the two loci used in the simulation are listed in Table 2. We repeated the simulation 100 times and analyzed the simulated data using two methods: (1) our Bayesian partition method using parallel tempering [18] with 15 temperature ladders, referred to as BP; (2) the two-stage regression method of Storey et al [5], referred to as SR. Details of the simulation and implementation of these two methods are described in the Supplemental Material. As shown from the receiver operating characteristic (ROC) curves in...
Figure 2. Comparison of different models for associating genotypes and phenotypes. (A) the regression model; (B) the Bayesian partition (BP) model with a single biallelic marker; (C) the BP model with two interactive biallelic markers. In the regression model, gene expression values (Y) are regressed onto marker genotypes (X). If the marker has a dominant effect on the gene expression, the regression implicitly partitions the expression values into low and high groups corresponding to genotypes aa for, say, low expression and Aa or AA for high expression. In the BP model, a latent variable, denoted here as “Individual Type”, is introduced and conditional on this variable the gene expression traits and marker genotypes are independent. In the case of a single marker, two individual types exist, colored here as green and red. In (c), gene expression is linked with a set of two biallelic markers. In this instance, individuals are partitioned into three types, colored here as green, blue and red, corresponding to low, medium, and high expression levels, respectively.

Figure 3, BP achieved a significantly higher power to detect eQTLs compared to SR. For example, allowing for 50 false positives, BP detected more than 500 (out of 640) true gene-expression and genotypes for 112 segregants from a cross between laboratory (BY) and wild (RM) strains of S. cerevisiae and detected 29 modules of genes and their associated markers (Methods). Among these 29 modules, 20 are linked to a single eQTL while the remaining nine are linked to two eQTLs. Three of the nine linking to two eQTLs give rise to significant epistatic interactions between the two loci. Twenty-six of the 29 modules significantly overlap (corrected p-value<0.05) with at least one of the average percentage of genetic variance explained by epistasis. In all modules, the heritability of the “core gene” is 0.6 and the average correlation of the module genes with the “core gene” is 0.5.

doi:10.1371/journal.pcbi.1000642.g002

doi:10.1371/journal.pcbi.1000642.t001

Yeast eQTL modules – a re-examination of the landscape of genetic complexity

We applied our Bayesian method to a data set consisting of gene expression and genotypes for 112 segregants from a cross between laboratory (BY) and wild (RM) strains of S. cerevisiae and detected 29 modules of genes and their associated markers (Methods). Among these 29 modules, 20 are linked to a single eQTL while the remaining nine are linked to two eQTLs. Three of the nine linking to two eQTLs give rise to significant epistatic interactions between the two loci. Twenty-six of the 29 modules significantly overlap (corrected p-value<0.05) with at least one of the eQTLs.
the 13 gene groups previously reported as mapping to eQTL hot spots [11]. We also tested each of these modules for enrichment using GO terms, a yeast knockout compendium [19], and transcription factor binding sites [20]. At p-value < 0.05 after multiple testing correction, 21 modules have at least one GO term enrichment; 22 modules overlap with at least one knockout signature, and 13 modules are enriched for at least one transcription factor binding site. The result is summarized in Table 3 and a breakdown result is in Supplementary Table S3. In contrast, the LOD score distributions of transcripts at the associated markers under the “single-transcript-single-marker” model are shown in Supplementary Figure S2. Our Bayesian method identifies significantly more weak gene-marker associations than the simple model. These GO enrichments support the biological relevance of different modules detected by our method. Each module is described in detail in the Supplemental Material

**Table 2. Genotypic means and frequencies for a two-locus model used in the simulation studies.**

<table>
<thead>
<tr>
<th>Locus 1</th>
<th>A</th>
<th>B</th>
<th>a</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Locus 2</td>
<td>p_{AB}</td>
<td>p_{AC}</td>
<td>p_{a}</td>
<td>p_{AB}</td>
</tr>
<tr>
<td>B</td>
<td>p_{AB}</td>
<td>p_{AC}</td>
<td>p_{a}</td>
<td>p_{AB}</td>
</tr>
<tr>
<td>b</td>
<td>p_{AB}</td>
<td>p_{AC}</td>
<td>p_{a}</td>
<td>p_{AB}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Locus 1</th>
<th>A</th>
<th>B</th>
<th>a</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Locus 2</td>
<td>p_{AB}</td>
<td>p_{AC}</td>
<td>p_{a}</td>
<td>p_{AB}</td>
</tr>
<tr>
<td>B</td>
<td>p_{AB}</td>
<td>p_{AC}</td>
<td>p_{a}</td>
<td>p_{AB}</td>
</tr>
<tr>
<td>b</td>
<td>p_{AB}</td>
<td>p_{AC}</td>
<td>p_{a}</td>
<td>p_{AB}</td>
</tr>
</tbody>
</table>

**Figure 3. Comparison of the receiver operator characteristic (ROC) curves for the gene-marker pair detection obtained by our Bayesian partition method (BP) and the two-stage regression method (SR).** Different points along the ROC curves represent the false positive and true positive counts averaged over 100 simulations at different posterior probability thresholds (for BP) or at different FDR thresholds (for SR). There are 40 genes in each of the eight modules which are linked to two markers and thus the total number of the true positive gene-marker pairs is 640.

doi:10.1371/journal.pcbi.1000642.g003

(0.731) is higher than that for modules 27 (0.409) and 28 (0.459). PHM7 was previously identified and validated as a causal regulator for this hot spot [10]. The PHM7 knockout signature significantly overlaps with modules 26 and 28 (p-value = 8.93 x 10^-5 and 0.0016, respectively). When compared to a previously constructed yeast knockout compendium [19], module 26 overlaps with 33 knockout signatures, while module 28 overlaps with only four of the knockout signatures (three of the four also overlap with module 26). Application of a causality test procedure [21] revealed that 52 genes (out of 83) in module 26 were supported as causal for at least one gene in module 28, while only six genes (out of 74) in module 28 were supported as causal for at least one gene in module 26 (shown in Supplementary Figure S3). These results indicate that genes in module 26 serve as the primary response to the causal perturbation of PHM7 and genes in module 28 serve as the secondary response. Other causal regulators for module 27 are independent of PHM7 may exist.

**Modules linked to two loci.** Our results provide a number of positive controls that illustrate how our method can detect complex eQTL hot spots into different modules and detect modules with complex genetic regulation. As summarized above, nine of the 29 modules we identified are linked to two eQTLs. Modules 3, 12 and 16 have significant epistatic interactions (p-value = 6.63 x 10^-5, 2.05 x 10^-13 and 0.00097, for the interaction terms, respectively) between the two loci. Modules 12 and 20 were previously reported in the literature [16–17]. Among the other seven novel modules, three of them (modules 16, 17 and 19) are likely due to cross-hybridization (see details in Supplemental Material Text S1). Module 3, which consists of many daughter cell expressed genes and is linked to two eQTLs with a significant epistatic interaction, is predicted to be under the regulation of AMN1 and BPH1, each located near the two eQTL loci. Modules 7 and 18 are each mapped to two previously detected eQTL hot spots suggesting that genes in these two modules are under the control of multiple mega-regulators.

The interaction term for module 12 is statistically most significant (p-value = 2.05 x 10^-13). A previous study [16] experimentally validated that an interaction between MAT at the chromosome III locus and GPA1 at the chromosome VIII locus affects a group of 19 genes. Among these 19 genes, one of them is not in our study set; two other genes were later experimentally verified to be “false positives” [16] and are correctly assigned to the null module in our analysis; and four other genes are negatively correlated with genes in this module and so are not placed in module 12. The remaining 12 genes are all recovered in this module. In addition, our method detects another gene, HMLALPHA2, which is also related to mating type. The heat map of the gene expression in module 12 is plotted in Figure 5B. This result demonstrates that our method not only is able to detect an experimentally validated interaction, but also has a higher specificity and sensitivity to detect the interaction than the regression based method.

Module 20 consists of 21 genes and is linked to two loci on chromosome XIII and X, respectively, but no epistatic interaction is detected between these loci. The heat map of the gene expression in this module is plotted in Figure 5C. Two transcription factor binding sites are enriched in the module, with the ZAPI binding site being the most significantly enriched (p-value = 9.51 x 10^-8). In fact, 14 of the 21 genes in module 20 are known or predicted to be ZAPI target genes [22] (p-value = 2.99 x 10^-22). ZAPI, which is physically located at the chromosome X locus and has an eQTL at the chromosome XIII locus, is included in this module. A previously identified ZAPI module [17] overlaps significantly with module 20. Among the ten genes in the ZAPI module, eight of them are also predicted in module 20. It was previously conjectured that a regulator at the
We have developed a Bayesian partition model for simultaneously mapping multiple eQTLs for multiple sets of co-regulated genes. Whereas conventional linkage analysis has been widely and successfully applied to the study of one or a small number of traits at a time, our module-based method is suitable for analyzing thousands of phenotypes simultaneously. Both simulation studies and empirical data examples demonstrated that our method is effective for detecting marker interactions, even when no marginal effects could be detected. These improvements in power are a direct result of accounting for the correlation among gene expression traits and assessing the joint effect of multiple eQTLs, including interactions, on these correlated gene sets.

One of the main advances in our approach is the introduction of the “individual type” as a latent variable to describe associations between gene expression traits and markers. The individual type latent variable can be interpreted as a classification of individuals according to a combination of phenotypes and genotypes. The underlying mathematical model for this dependence structure is represented as a chain in which the joint distribution for some set of markers influences a set of expression traits via a latent “individual type” variable. After integrating out this latent variable, we observe a direct relationship between the marker and gene expression sets, similar to what would have been obtained from a traditional regression model in the single-marker, single-gene case (Figures 2A and 2B). However, the advantage over the standard regression in introducing the latent variable can be interpreted as a classification of individuals according to a combination of phenotypes and genotypes.

Figure 4. The posterior probability plot. The height of each point is the posterior probability for the most likely classification of the gene; and the shape/color of the point represents the module type of the classification. The first 200 genes are those in one of the four non-null modules, separated by vertical lines.

doi:10.1371/journal.pcbi.1000642.g004
<table>
<thead>
<tr>
<th>Module</th>
<th>Size</th>
<th>Loci</th>
<th>GO category</th>
<th>KO</th>
<th>TFBS</th>
<th>eQTL hot spot</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>38</td>
<td>Chr II: 548401</td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>33</td>
<td>Chr II: 548401</td>
<td></td>
<td>1</td>
<td>0</td>
<td>2****</td>
</tr>
<tr>
<td>3</td>
<td>16</td>
<td>Chr II: 548401</td>
<td>cell wall (sensu Fungi)**</td>
<td>6</td>
<td>3</td>
<td>2**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chr III: 177850</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>137</td>
<td>Chr II: 548401</td>
<td>Nucleolus****</td>
<td>9</td>
<td>0?</td>
<td>2**</td>
</tr>
<tr>
<td>5</td>
<td>75</td>
<td>Chr II: 548401</td>
<td></td>
<td>0</td>
<td>1</td>
<td>2**</td>
</tr>
<tr>
<td>6</td>
<td>38</td>
<td>Chr II: 620212</td>
<td>Protein disulfide isomerase activity**</td>
<td>2</td>
<td>1</td>
<td>8**</td>
</tr>
<tr>
<td>7</td>
<td>83</td>
<td>Chr III: 79091</td>
<td>'de novo' IMP biosynthesis**</td>
<td>17</td>
<td>2</td>
<td>4 ****</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chr XV: 170945</td>
<td></td>
<td></td>
<td></td>
<td>10**</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>12***</td>
</tr>
<tr>
<td>8</td>
<td>69</td>
<td>Chr III: 79091</td>
<td>histidine biosynthesis*</td>
<td>53</td>
<td>2</td>
<td>4****</td>
</tr>
<tr>
<td>9</td>
<td>61</td>
<td>Chr III: 79091</td>
<td></td>
<td>7</td>
<td>0</td>
<td>4**</td>
</tr>
<tr>
<td>10</td>
<td>18</td>
<td>Chr III: 81832</td>
<td>branched chain family amino acid biosynthesis*</td>
<td>18</td>
<td>1</td>
<td>4****</td>
</tr>
<tr>
<td>11</td>
<td>52</td>
<td>Chr III: 81832</td>
<td>nuclear nucleosome***</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chr VIII: 84437</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>13</td>
<td>Chr III: 201166</td>
<td>Regulation of transcription from RNA polymerase II promoter*</td>
<td>1</td>
<td>0</td>
<td>4****</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chr VIII: 111679</td>
<td></td>
<td></td>
<td></td>
<td>5*</td>
</tr>
<tr>
<td>13</td>
<td>9</td>
<td>Chr III: 201166</td>
<td></td>
<td>10</td>
<td>3</td>
<td>4****</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5**</td>
</tr>
<tr>
<td>14</td>
<td>13</td>
<td>Chr V: 116530</td>
<td>'de novo' pyrimidine base biosynthesis**</td>
<td>4</td>
<td>0</td>
<td>6****</td>
</tr>
<tr>
<td>15</td>
<td>44</td>
<td>Chr VIII: 111690</td>
<td>Mating projection tip***</td>
<td>20</td>
<td>3</td>
<td>7****</td>
</tr>
<tr>
<td>16</td>
<td>10</td>
<td>Chr X: 22315</td>
<td>aldehyde metabolism***</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chr VI: 28041</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>11</td>
<td>Chr XII: 659357</td>
<td></td>
<td>12</td>
<td>0</td>
<td>8**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chr XIII: 430164</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>45</td>
<td>Chr XII: 662627</td>
<td>ergosterol biosynthesis****</td>
<td>6</td>
<td>1</td>
<td>8****</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chr III: 79091</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>34</td>
<td>Chr XII: 105609</td>
<td>telomerase-independent telomere maintenance***</td>
<td>11</td>
<td>0</td>
<td>9****</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chr IV: 1525327</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>21</td>
<td>Chr XIII: 49903</td>
<td></td>
<td>4</td>
<td>2</td>
<td>10***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chr X: 327852</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>81</td>
<td>Chr XIV: 449639</td>
<td>endoplasmic reticulum***</td>
<td>2</td>
<td>0</td>
<td>1*</td>
</tr>
<tr>
<td>22</td>
<td>52</td>
<td>Chr XIV: 486861</td>
<td>structural constituent of ribosome****</td>
<td>2</td>
<td>0</td>
<td>11****</td>
</tr>
<tr>
<td>23</td>
<td>68</td>
<td>Chr XIV: 486861</td>
<td>Arp2/3 protein complex**</td>
<td>0</td>
<td>0</td>
<td>11**</td>
</tr>
<tr>
<td>24</td>
<td>39</td>
<td>Chr XIV: 449639</td>
<td>nuclear pore organization and biogenesis*</td>
<td>0</td>
<td>0</td>
<td>11****</td>
</tr>
<tr>
<td>25</td>
<td>77</td>
<td>Chr XIV: 486861</td>
<td>mitochondrial inner membrane**</td>
<td>0</td>
<td>0</td>
<td>11****</td>
</tr>
<tr>
<td>26</td>
<td>83</td>
<td>Chr XV: 170945</td>
<td>response to stress***</td>
<td>33</td>
<td>1</td>
<td>12****</td>
</tr>
<tr>
<td>27</td>
<td>45</td>
<td>Chr XV: 170945</td>
<td></td>
<td>0</td>
<td>0</td>
<td>12****</td>
</tr>
<tr>
<td>28</td>
<td>74</td>
<td>Chr XV: 170945</td>
<td>Fructose transporter activity*</td>
<td>4</td>
<td>0</td>
<td>12****</td>
</tr>
<tr>
<td>29</td>
<td>42</td>
<td>Chr XV: 563943</td>
<td>respiratory chain complex III (sensu Eukaryota)****</td>
<td>10</td>
<td>5</td>
<td>13****</td>
</tr>
</tbody>
</table>

*aNumber of genes in each module.
*bThe chromosome positions of markers associated with each module.
*cThe most significant GO terms. A total of 510 GO terms of sizes 5 to 300 were tested. Multiple testing corrected (Fisher Exact Test p-value × 510) p-values less than 0.05 are displayed at four different levels indicated by *, **: 10^-5~0.05; ***: 10^-5~10^-10; ****: 10^-10~10^-15.
*dNumber of knockout signatures that overlap with each module. 287 knockout signatures [19] were tested and the p-value cut-off is 1:74×10^-4 (0.05/287).
*eNumber of the transcription factors whose binding sites are enriched in each module. 119 transcription factor binding sites [20] were tested and the p-value cut-off is 4.2×10^-4 (0.05/119).
*fOverlapped eQTL hot spots. Multiple testing corrected (Fisher Exact Test p-value × 13) p-values at cut-off 0.05 are displayed in four different levels indicated by *.

Module 4 is enriched with de novo motifs PAC and RRPE.

doi:10.1371/journal.pcbi.1000642.t003
individual type variable is its enabling us to model epistatic interactions and pleiotropy simultaneously.

Linkage disequilibrium (LD) among adjacent markers is an important feature of the genetic marker data. For individuals produced by the laboratory crosses (e.g., F1 and F2 designs), the marker dependency can be modeled satisfactorily by a Markov chain. The BP model can easily entertain this modification of the background marker distribution, but the computation time required to run this modified model dramatically increases since we need a forward-summation-backward-sampling algorithm to update the marker indicators (see Supplemental Material Text S1 for details).

Another ad hoc strategy to account for the marker correlations without directly modeling them was to first scan all markers and to enumerate those marker pairs with correlations exceeding a given threshold. Then, in the MCMC algorithm, we imposed a mutually exclusive condition for such pairs so that highly correlated marker pairs would not appear simultaneously in any module.

We compared the Markov model approach with the ad hoc strategy on a small simulated data sets and a subset of the real data (data not shown). The ad hoc strategy always provided nearly identical results to that of the Markov model with only a fraction of the computation cost. Note that there are also markers that are highly correlated but are not physically linked [26]. In such cases the Markov model actually worked less satisfactorily than the ad hoc approach.

Our method shares some similarities to other methods in the literature, but also shows clear distinctions. For example, Lee et al. [17] proposed to simultaneously partition the gene expression and genotype markers. However, their method requires strong priors on the potential regulators, while our method does not. Kendziozki et al. [14] proposed a mixture of markers model to find the eQTLs for multiple gene expression. However, their method separates the gene clustering and eQTL mapping steps, where they first use k-means clustering to identify subsets of genes, and then apply eQTL mapping to the clusters of genes. In addition, their method does not address the epistatic effects. In contrast, gene expression partition and eQTL mapping are modeled jointly in our Bayesian method, and we are able to effectively detect epistasis by using a comprehensive statistical model on both the gene expression and the markers. Our analysis of the yeast data identified 20 modules linked to one eQTL and 9 modules linked to two eQTLs, among which three giving rise to strong epistatic interactions between markers. Some of the modules linked to two eQTLs are consistent with previously reported results [5,17], and we were able to identify more true positive hits along with fewer false positives than previously reported.

Figure 5. Heat map for expression of genes in modules. (A) for module 3; (B) for module 12; (C) for module 20. Each row represents a gene with the gene name listed on the right and each column represents an individual. Individuals are divided into four groups according to the genotypes of the two eQTLs. Over- and under-expression are indicated by red and green, respectively.

doi:10.1371/journal.pcbi.1000642.g005
A Bayesian Partition Method

Methods

Bayesian partition model

A module is defined in the Results section as a set of gene expression traits (referred to as “genes” henceforth) and a set of genetic markers (e.g., SNPs) such that the mRNA expression variation of the genes is associated with the allelic variation of the markers. This association between multiple genes and markers is characterized by a latent indicator variable, individual type, conditional on which the trait and marker variables are independent of each other. The individual type latent variable can be viewed as representing a certain combination of markers that induces changes in expressions of a certain set of genes across different individual types.

To formally describe our model, consider a sample with \( N \) individuals. Each individual \( i \) is measured with \( G \) gene expression values denoted as \( \{ y_{ig} : g = 1, \ldots, G \} \) and \( M \) marker genotypes denoted as \( \{ x_{im} : m = 1, \ldots, M \} \). We assume that the observed data can be partitioned into \( D \) nontrivial modules plus a null component. The number of non-null modules, \( D \), is pre-specified by the user and should reflect the user’s prior belief in the higher level structure of the data. Every gene \( g \) or marker \( m \) belongs to one of the \( D \) nontrivial modules or the null module, determined by the gene indicator \( I_g \in \{0,1,\ldots,D\} \) and the marker indicator \( J_m \in \{0,1,\ldots,D\} \). For each module \( d \in \{1,\ldots,D\} \), we further partition the \( N \) individuals into \( n^d \) types denoted by the individual indicators \( K_{id} \in \{1,\ldots,n^d\} \) for \( i \in \{1,\ldots,N\} \). Each module may have a different number of individual types as well as different ways of partitioning the \( N \) individuals. For example, with a single biallelic marker (alleles ‘A’ and ‘a’) in the module, the module may have two individual types corresponding to genotypes aa vs. Aa or AA (dominant model), or 3 individual types corresponding to genotypes aa, Aa and AA (additive model). We seek module partitions in which expression patterns are similar for all genes, and gene expression variations across different individuals can be explained by the individual types. A cartoon illustration of the partition model is shown in Figure 1.

We model the gene expression traits in module \( d \) by an ANOVA model so that each trait value is the sum of the gene effect (\( \beta_g \)), the eQTL effect for individual type \( k (\beta_k) \), the individual effect (\( r_i \)), and an error term:

\[
y_{ig} = \beta_g + \beta_k + r_i + e_{ig},
\]

where gene \( g \) is in module \( d \), \( k \) is the individual type of \( i \), and \( r_i \) and \( e_{ig} \) are random effects, following independent Gaussian distributions with mean zero.

To account for epistasis, we model the joint distribution of all the associated markers of module \( d \), \( \mathbf{x}_i = \{ x_{im} : m \text{ is in module } d, \text{ i.e., } J_m=d \} \), by a multinomial distribution, whose frequency vector is determined by the individual type \( k \), i.e.,

\[
\mathbf{x}_i \sim \text{Multinomial}(1; \mathbf{\hat{\beta}}_k).
\]

For example, if there are two markers \( \{m_1, m_2\} \) in the module and each has three genotypes, then there are nine combinations of the marker patterns. Thus \( \{x_{im_1}, x_{im_2}\} \) follows a 9-dimensional multinomial distribution.

For the null component, we assume that there is no association between the genes and the markers. The gene expression traits follow a normal distribution and the marker genotypes follow an independent multinomial distribution.

To avoid overfitting, we put an exponential prior on the indicator variables to penalize partitions with high complexity:

\[
P(\mathbf{I}, \mathbf{J}_m, \mathbf{K}_d) \propto \exp \left( -c_G \sum_d n^d G - c_M \sum_d L^d T - c_T \sum_d n^d T \right),
\]

where \( n^d G, n^d M, n^d T \) are the number of genes, markers and individual types in module \( d \), and \( L \) is the number of genotypes at each marker. We use conjugate priors on the continuous parameters, such as means and variances of the Gaussian distributions and frequency vectors of the multinomials, so that most of these...
parameters can be integrated out analytically to reduce the complexity of the posterior distribution.

The joint posterior distribution of all unknown variables is of the form:

$$P(\hat{I}_t, \hat{J}_m, \hat{K}_b, \beta | X, Y) \propto P(\hat{I}_t, \hat{J}_m, \hat{K}_b) \times P(X, Y | \beta, \hat{I}_t, \hat{J}_m, \hat{K}_b),$$

where $\beta$ represents the set of left-over continuous parameters unable to be integrated out analytically. In order to make inference on the eQTL modules from this posterior distribution, we construct a Markov chain Monte Carlo method to traverse the joint space of all unknown parameters. Each Markov chain is randomly initialized, and uses the Gibbs sampler and the Metropolis-Hasting algorithm to update the variables. We implement a split-merge algorithm, which is a special case of the reversible jump MCMC [29], to update the individual partitions globally. Parallel tempering [30] is used to help mixing the Markov chain. Further details of the modeling and sampling strategies can be found in the Supplemental Material Text S1.

Posterior probabilities are evaluated for each gene and candidate marker set to belong to each module based on the Monte Carlo samples. A threshold is then applied to the posterior probabilities to determine whether a particular gene and marker set should be included in a module.

### Application to the yeast data set

We assembled genotypic and expression data from 112 segregants obtained from a previously described yeast cross between the BY and RM strains of *S. cerevisiae* [15]. Of the 5,740 genes represented on the microarrays in this study, we selected 3,662 informative genes as input into the partition algorithm following the same criteria as previously described [10]. We then transformed the gene expression values by first performing quantile normalization [31] to make the distribution of the log-expression ratios for each individual to be the same, and then normalizing each gene so that the mean expression level for each gene was 0 and the standard deviation was 1.

Given that genes in the data set have been previously mapped to 15 distinct eQTL hot spots [11] and that there can be multiple causal factors for a single eQTL hot spot, we set the number of starting modules for our MCMC algorithm to 35 (3×13 plus a null model) to account for these previously identified groups, and to also allow for the detection of new groups as well. For the parallel tempering implementation, we used 30 temperature ladders with almost equal spacing so that the average acceptance probability for exchanges between adjacent chains was roughly 0.15–0.3. We ran MCMC sampling for 1,000,000 iterations in each chain, which took one week of 30 CPUs (accounting for 30 parallel temperature ladders of the MCMC algorithm) on a Linux cluster with 2GHz CPUs. The log posterior probability and its auto-correlation curve depicted in Figures S5C and S5D highlight that the Markov chain became stationary after a burn-in period. See Supplemental Material Text S1 for more details.

Because markers in the yeast data set are very densely distributed, adjacent markers are almost always highly correlated. After MCMC sampling, markers adjacent to the “truly” linked marker often diluted the posterior probability for the true marker-module linkage. Since a proper Markov chain model for unlinked markers is computationally too expensive to implement (see Supplemental Material Text S1), we employed a heuristic method to counter this problem. We first specified a window centered at each marker so that markers inside the window are in high LD with the marker at the center. The posterior probabilities of all markers in the window were summed up and regarded as the modified posterior probability for the central marker. The markers with peak probabilities exceeding the given threshold were selected and all other markers in the corresponding windows were masked out.

Although we did not explicitly model pleiotropic effects for markers (i.e., single markers were not allowed to be associated with expression traits in multiple modules), we reported several modules mapped to the same markers in the yeast data set (See Table 3 and discussions in the Supplemental Material Text S1). The reason for this apparent contradiction is due to the aforementioned moving window approach and the dense distribution of the markers. In other words, if marker $m$ is truly linked to two modules, in computation its adjacent markers can serve as its surrogates so that a subset of these markers are mapped to module 1, and the remainders mapped to module 2. Then the use of the moving window method can restore the total probability back to marker $m$.

To test the robustness of our result with respect to the initial parameters, we ran our program using three different numbers of modules, $D = 35$, $D = 40$, and $D = 45$, each having three independent runs. Samples from the run with the highest average posterior probability for each value of $D$ were used in the subsequent analyses. We chose 0.8 as the threshold for the posterior probabilities to determine the module membership for each gene and marker. We observed that more than 70% of the genes were consistently grouped together and mapped to the same markers (or null module) in all the runs with different $D$ values. These genes and their associated markers formed the list of 29 modules.

### Supporting Information

**Text S1** Supplementary methods and results

Found at: doi:10.1371/journal.pcbi.1000642.s001 (0.49 MB PDF)

**Figure S1** Module-by-module comparison of the Bayesian partition (BP) method and the step-wise regression (SR) method. (A) Number of the true positive gene-marker pairs detected in each module by the BP method (top) and the SR method (bottom). Nine different lines correspond to different posterior probability thresholds (for BP) or different FDR thresholds (for SR), both of which decrease from 0.9 to 0.1 linearly. There are 40 genes in each of the eight modules which are linked to two markers and thus the number of the true positive gene-marker pairs is 640. (B) Barplots of the number of true eQTLs detected in each module by the BP method (blue) and SR method (green). The shaded bar represents the number of genes detected as mapped to at least one of the true eQTLs while the solid bar represents the number of genes detected as mapped to both eQTLs. The thresholds are 0.5 for both posterior probability (BP) and FDR (SR). From Figure 1 we know that the total number of false positive gene-marker pairs is 11.41 and 38.04 for BP and SR respectively. When the thresholds are relaxed to 0.1, more eQTLs were detected in each category, as indicated by the vertical lines above the bars. However, the total number of the false positive gene-marker pairs is still lower using BP (178.37) compared to that using SR (267.07). Found at: doi:10.1371/journal.pcbi.1000642.s002 (0.36 MB TIF)

**Figure S2** The distributions of LOD scores under the “single-gene-single-marker” model for genes in the 29 modules identified by the Bayesian method. (A) The LOD score distribution for genes in modules linked to a single eQTL. The LOD scores for 56.3% of transcripts were less than 4.35, the threshold corresponding to a genome-wide FDR of 0.01, and 11.5% of transcripts were less than 1.45, corresponding to a point-wise FDR of 0.01. (B) the LOD score distribution for genes in modules linked to two eQTLs.
The LOD scores for 69% and 32.5% of transcripts were less than 4.35 and 1.45, corresponding to a genome-wide and a point-wise FDR of 0.01, respectively.

Found at: doi:10.1371/journal.pcbi.1000642.s003 (0.11 MB TIF)

**Figure S3** Plot of the causality test results for all pairs of genes between (A) module 4 and module 5 and (B) module 26 and 29. For a particular pair of genes (G1, G2) from module 4 and module 5, respectively, if the causality test claims that gene G1 is causal to gene G2 (corrected p-value<0.05), i.e. G1→G2, then a green dot is plotted at the corresponding position. Similarly, if the causality test results in G2→G1, then a red dot is plotted at the corresponding position. Genes in module 4 and module 5 are sorted for better visualization. Similar procedure applies to (B).

Found at: doi:10.1371/journal.pcbi.1000642.s004 (0.34 MB TIF)

**Figure S4** A local view of the coding sequence alignment of RM vs. BY for gene BPH1. The RM sequence has a deletion in the position labeled in red which results in an in-frame stop.

**Figure S5** Trace plots and autocorrelation plots of the log posterior probabilities for one of the simulated data set ((A) and (B)) and the yeast data set analysis ((C) and (D)). In (A), the trace plot was generated from two independent chains, each having 100,000 iterations, and the autocorrelation plot in (B) was obtained from the first chain at every 50 iterations. In (C), trace plot was generated from 1,000,000 Markov chain iterations, using D = 40 (D is the number of the modules). The last 700,000 iterations were used to generate the auto-correlation plot in (D).

Found at: doi:10.1371/journal.pcbi.1000642.s005 (0.03 MB TIF)

**Table S1** Design for the simulation II.

Found at: doi:10.1371/journal.pcbi.1000642.s007 (0.20 MB PDF)

**Table S2** True markers and inferred markers in each module.

Found at: doi:10.1371/journal.pcbi.1000642.s006 (0.14 MB PDF)

**Table S3** Enrichment of (A) gene knockout signatures and (B) TFBS for each module.

Found at: doi:10.1371/journal.pcbi.1000642.s009 (0.15 MB PDF)

**Author Contributions**

Conceived and designed the experiments: WZ JZ EES JSL. Analyzed the data: WZ JZ. Contributed reagents/materials/analysis tools: WZ JZ. Wrote the paper: WZ JZ. EES JSL.

## References