An efficient approach to large-scale genotype–phenotype association analyses

Runqing Yang, Hongwang Li, Lina Fu and Yongxin Liu

Submitted: 7th May 2013; Received (in revised form): 31st July 2013

Abstract

Modern molecular biotechnology generates a great deal of intermediate information, such as transcriptional and metabolic products in bridging DNA and complex traits. In genome-wide linkage analysis and genome-wide association study, regression analysis for large-scale correlated phenotypes is applied to map genes for those by-products that are regarded as quantitative traits. For a single trait, least absolute shrinkage and selection operator with coordinate descent step can be employed to efficiently shrink sparse non-zero genetic effects of quantitative trait loci (QTLs). However, regression analyses in a trait-by-trait basis do not take account of the correlations among the analyzed traits. In this study, conditional phenotype of each trait is defined, given other traits. Large-scale genotype–phenotype association analyses are therefore transformed to separate genotype-conditional phenotype ones. Meanwhile, the correlation architecture between each trait and other traits can also be provided by shrinkage estimation for each conditional phenotype. Simulation demonstrates that the proposed conditional mapping method is generally identical to joint mapping method based on multivariate analysis in terms of statistical detection power and parameter estimation. Application of the method is provided to locate eQTL in yeast.

Keywords: large-scale; genotype–phenotype association; conditional phenotype; LASSO; regression analysis

INTRODUCTION

Quantitative trait loci (QTL) mapping is essentially the statistical study of genotype–phenotype association, with the purpose to elucidate genomic variation for quantitative traits. The extension of phenotype has been expanded from economically important traits and/or complex diseases to thousands of transcriptional expressions. As intermediate phenotype to bridge between DNA and complex quantitative traits, proteomic and metabolomic products can also be regarded as endophenotypes [1]. With the development of high-throughput genotyping technology, millions of single nucleotide polymorphisms (SNPs) have been used in genome-wide association study (GWAS). Genotype–phenotype association analysis is prodigiously challenged to handle high-dimensional phenotype and genomic marker data.

For multiple correlated traits, the advantages of joint mapping method over separate mapping method have been well recognized. Statistically, joint analysis increases the power to detect QTL and precision of parameter estimation [2, 3]. Biologically, joint analysis is able to address the issue of pleiotropy versus the close linkage [4, 5] and to access the endophenotypes intermediate between a...
gene and a trait. The joint mapping method is usually based on multivariate regression analysis for multiple traits with normal distribution. Owing to the need of matrix calculation, however, its applicability is confined to the number of traits analyzed. Statistical dimension reduction technology, such as principal component analysis and discriminant analysis, is therefore introduced to map QTLs for many quantitative traits [6–10]. The mapping results from principal component analysis are theoretically interpretable with principal components as super traits, as QTL genetic effects on super traits can be back-transformed to original traits with principal eigenvector matrix. The problem of using principal component mapping is that the back-transformed least square and maximum likelihood estimates are not equivalent to those from joint mapping method [10–12]. Instead of using traditional multivariate regression model that assumes the same genetic model for all traits, seemingly unrelated regression model [13] is proposed to describe either the same or different QTLs for different traits, which improves the power to detect genetic architecture for multiple traits and facilitates distinguishing pleiotropic effect versus close linkage [14, 15]. Generalized estimating equations have been applied to multivariate QTL mapping for multiple discrete traits [16–18] as well as mixed normal and binary traits [19–21]. Additionally, Bayesian network [22] is a promising approach for multivariate mapping [23], which can integrate genetic epidemiological data with these types of flow networks to better understand the genotype–phenotype map [24–26].

QTL mapping for transcriptional expressions, i.e. eQTL mapping, is a typical genotype–phenotype association analysis for high-dimensional data [27–32]. The aforementioned mapping methods for multiple quantitative traits do not appear to be appropriate for mapping eQTLs due to large dimension with transcriptional expressions. By clustering transcripts with similar expression into groups, sparse partial least-squares regression method has been developed to select markers associated with each cluster of genes [33]. Adaptive multi-task least absolute shrinkage and selection operator (LASSO) [34, 35] has been developed for detecting eQTLs, which takes into account related expression traits simultaneously while incorporating many regulatory features. In contrast, the graph-guided fused LASSO [36, 37] can consider networks over multiple expression traits within an association analysis, without incorporating prior knowledge of genomic locations. No matter which multivariate mapping method is used, however, simultaneous analysis for highly-dimensional phenotypes and genotypes is infeasible under computational conditions available now. In this study, conditional phenotype of each trait is defined based on multivariate normal distribution of all analyzed traits. Extremely fast LASSO with coordinate descent step is adopted to separately estimate conditional phenotypes and sparse non-zero genetic effects of all loci over the entire genome on each conditional phenotype. Computer simulation is used to demonstrate statistical efficiency of the so-called conditional mapping method. A yeast data set is reanalyzed to map eQTLs.

METHODS

Genetic model for conditional analysis

For genotype–phenotype association analysis, markers are genotyped and phenotypes are observed on subjects. Subjects may be from a genetically designed population for linkage analysis, or from a natural population for GWAS. Assuming that total t traits and m markers are measured for n individuals, the relationship between phenotypes of those traits and markers are generally described in the form of multivariate linear model:

\[
\begin{align*}
  y_{il} &= \sum_{j=1}^{i} x_{ij} \beta_{lj} + \sum_{j=1}^{m} z_{ij} \gamma_{lj} + e_{il} \\
  y_{il} &= \sum_{j=1}^{i} x_{ij} \beta_{lj} + \sum_{j=1}^{m} z_{ij} \gamma_{lj} + e_{il} \\
  y_{il} &= \sum_{j=1}^{i} x_{ij} \beta_{lj} + \sum_{j=1}^{m} z_{ij} \gamma_{lj} + e_{il}
\end{align*}
\]

where \( \beta_{lj} \) is the jth systemic environmental effect for the \( i \)th trait, \( x_{ij} \) is an incidence variable for the \( i \)th subject in the \( j \)th systemic environment, \( \gamma_{lj} \) is the genetic effect of the \( j \)th marker on the \( i \)th trait, \( z_{ij} \) is an indicator variable of the \( j \)th marker for the \( i \)th subject, determined by the genotypes of the marker, \( e_{il} \) is the residual error, which is assumed to follow a multivariate normal distribution \( e_{il} \sim N_{t}(0, \Sigma_{e}) \) where \( \Sigma_{e} \) is the residual covariance matrix.

To map QTLs for multiple traits, one simplest approach is to analyze each trait separately, the so-called separate mapping by trait. Under the least squares framework, the univariate analysis can obtain identical estimates of genetic effect compared to multivariate analysis, but its test procedure does not consider correlations among traits. By calculating relative residual covariance matrices, multivariate
analysis infers significance of genetic effect $\gamma_b$ with the likelihood ratio statistics \cite{38}:
\[
\chi^2_y = -2n \ln \frac{\Sigma_y}{\tilde{\Sigma}_y}
\] (2)

which follows a Chi-square distribution with one-degree of freedom. In the likelihood ratio, $\Sigma_y$ is the residual covariance matrix for Equation (1) exclusive $\gamma_y$. The QTL mapping with such a statistic is defined as joint mapping, as it uses statistical information from joint analysis of multiple traits. Even with few traits analyzed, large computing burden can incur for calculating the determinant of matrix $\Sigma_y$. When many traits are analyzed, calculating the determinants of residual covariance matrices with high dimension can cause the joint mapping fail to work.

In the notation of matrix, Equation (1) can be written as
\[
y_i = \sum_{j=1}^{n} x_{ij} \beta_j + \sum_{j=1}^{m} z_{ij} \gamma_j + \epsilon_i
\] (3)

with $y_i = [y_i \ y_j \ \cdots \ y_t]^T$, $\beta_j = [\beta_{1j} \ \beta_{2j} \ \cdots \ \beta_{nj}]^T$ and $\gamma_j = [\gamma_{1j} \ \gamma_{2j} \ \cdots \ \gamma_{nj}]^T$.

The expectation of $y_i$ is
\[
E(y_i) = \mu_i = \sum_{j=1}^{n} x_{ij} \bar{\beta}_j + \sum_{j=1}^{m} z_{ij} \bar{\gamma}_j
\] (4)

and its covariance matrix $V(y_i) = \Sigma_y$.

Let $y_{il} = [y_i \ y_{i(l-h)}]$ where $y_{i(l-h)}$ is a vector representing all of the elements except $y_i$. Similarly, $\mu_i$ and $\Sigma_e$ can be partitioned in the same manner as
\[
\mu_i = \begin{bmatrix} \mu_i, \mu_{i(l-h)} \end{bmatrix} \\
\Sigma_e = \begin{bmatrix} \Sigma_{e(l-h)} & \Sigma_{e(l-h)} \\ \Sigma_{e(l-h)} & \Sigma_{e(l-h)} \end{bmatrix}
\]

Following the definition of conditional random variables under multivariate normal distribution, $y_{il}$ conditional on $y_{i(l-h)}$, denoted by $y^*_{il}$, can be expressed as
\[
y^*_{il} = y_{il} | y_{i(l-h)} = y_{il} - \Sigma_e y_{i(l-h)(-l)} (y_{i(-l)} - \mu_{i(l-h)})
\] (5)

It is easy to prove that $y^*_{il}$ for $l = 1, 2, \cdots, t$ is independent of $y_{i(l-h)}$ in statistics.

Hence, the genetic model for $y^*_{il}$ is constructed as
\[
y^*_{il} = \mu_i + \epsilon_i = \sum_{j=1}^{n} x_{ij} \beta_j + \sum_{j=1}^{m} z_{ij} \gamma_j + \epsilon_i
\] (6)

which is applied to $l = 1, 2, \cdots, t$; residual error $\epsilon_i \sim N(0, \Sigma_{e(l-h)} - \Sigma_{e(l-h)} \Sigma_{e(-l-h)}^{-1} \Sigma_{e(-l-h)})$, which is also independent of the residuals for the remaining $t - 1$ traits analyzed in Equation (1). Equation (6) can be therefore used for separate QTL mapping trait by trait, as statistical inference for such a model is not influenced by other traits. Herein, this mapping method is called conditional mapping.

### Statistical inference for genetic effects

As a linear model, Equation (6) may not be straightforwardly solved by ordinary least square method because the dependent variable $y^*_{il}$ is unknown and the number of genetic effects is generally greater than that of subjects in the practice of QTL mapping. Moreover, there is normally small number of non-zero genetic effects in Equation (6). Given conditional phenotype $y^*_{il}$, the LASSO with a coordinate descent step \cite{39, 40} can efficiently shrink most of genetic effects to zero by solving the objective function:
\[
\min \left[ -\frac{1}{2n} \sum_{i=1}^{n} \left( y^*_{il} - \sum_{j=1}^{n} x_{ij} \beta_j \\
- \sum_{j=1}^{m} z_{ij} \gamma_j \right)^2 + \lambda_1 \sum_{i=1}^{n} |\gamma_i| \right]
\] (7)

where, $\lambda_1$ is a tuning parameter, which will be chosen by cross-validation.

Owing to unknown $y^*_{il}$, $\Sigma_{e(l-h)} \Sigma_{e(-l-h)}^{-1}$ amounts to the regression coefficients of the residual $y_{il} - \mu_l$ on the residuals $(y_{i(l-h)} - \mu_{i(l-h)})$, which are estimated by the linear model:
\[
y_{il} - \mu_l = \sum_{j=1, j \neq l}^{t} \beta_j (y_{ij} - \mu_j) + \epsilon_{il}
\] (8)

If large-scale traits are analyzed, the number of estimated parameters in Equation (8) will be far greater than that of subjects. Non-zero regression coefficients in Equation (8) are obtained with shrinkage estimation as well, which involves solving
\[
\min \left[ -\frac{1}{2n} \sum_{i=1}^{n} \left( y_{il} - \mu_l - \sum_{j=1, j \neq l}^{t} \beta_j (y_{ij} - \mu_j) \right)^2 + \lambda_2 \sum_{j=1, j \neq l}^{t} |\beta_j| \right]
\] (9)

where, $\lambda_2$ is also a tuning parameter.

Then, $y^*_{il}$ is calculated by $y_{il} - \sum_{j=1, j \neq l}^{t} \hat{\beta}_j (y_{ij} - \mu_j)$. Most importantly, the solution for $\Sigma_{e(-l-h)}^{-1}$ can be avoided here.

According to the equivalency of least square or maximum likelihood estimates for genetic effects
between univariate analysis and multivariate analysis, the expectation \( \mathbf{\mu} \), including systemic environmental effects and genetic effects in Equation (1) can be estimated by using the LASSO trait by trait, as follows

\[
\beta_j = \text{arg min} \left[ -\frac{1}{2n} \sum_{i=1}^{n} \left( y_{il} - \sum_{j=1}^{m} x_{ij} \beta_j \right)^2 + \lambda_3 \sum_{j=1}^{m} |\gamma_j| \right].
\]

(10)

with \( \lambda_3 \) being a tuning parameter.

Based on the solution strategy described earlier in the text, parameter estimation for Equation (6) can be summarized as

1. Estimate the expectation \( \mathbf{\mu}_{il} \) for each trait by solving equation (10).
2. Calculate conditional phenotype \( y^*_{il} \) for each trait by solving equation (9).
3. Estimate non-zero genetic effects for each conditional trait by solving equation (7).

Non-zero genetic effects obtained from step (3) are basically the same as those from step (1), at least for the loci corresponding to non-zero genetic effects. Therefore, step (3) can be omitted and directly moved to statistical inference step for genetic effects. After shrinkage estimation for genetic effects, few non-zero genetic effects are retained in Equation (6). The number of non-zero effects is typically less than the sample size; hence, ordinary least squares can then be used to estimate the non-zero genetic effects and statistically infer their significance. The loci or genetic markers corresponding to significant genetic effects are determined as the QTLs for those phenotypes.

### RESULTS

**Simulation study**

The objective of simulations is to demonstrate statistical efficiency of the conditional mapping method proposed here by comparing it with separate mapping method and joint mapping method. In a backcross population, 1200 co-dominant markers are evenly distributed on six chromosomes and 200 markers are equidistantly placed on each chromosome and each chromosome is 200 cM long. Two, two, one, two, one and two QTLs are separately put on six chromosomes, whose positions and genetic effects are listed in Tables 1 and 2. Only two traits are simulated, whose phenotypic values can be drawn from bivariate normal distribution with the expectation \( \mathbf{\mu} \) and residual covariance matrix \( \Sigma_2 \). After the QTL genotypes are simulated along with marker genotypes, the expectation \( \mathbf{\mu} \), without systemic environmental effect, is calculated by multiplying the indicator variable corresponding to QTL genotypes and the genetic effects. Assuming residual variance of one, the residual covariance depends on the correlation between the two traits. By taking indicator variables of QTL genotypes as +1 for heterozygote and −1 for homozygote, QTL heritability is estimated as the ratio of square of QTL genetic effect to phenotypic variance, where phenotypic variance is sum of squares for genetic effects of all simulated

### Table 1: Mean estimates and standard deviations (in parentheses) of QTL positions detected with three mapping methods for the simulated data sets with correlation 0.6

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Method</th>
<th>Chr. no.</th>
<th>QTL no.</th>
<th>C_1</th>
<th>C_2</th>
<th>C_3</th>
<th>C_4</th>
<th>C_5</th>
<th>C_6</th>
</tr>
</thead>
<tbody>
<tr>
<td>150</td>
<td>True position</td>
<td>46</td>
<td>112</td>
<td>96</td>
<td>186</td>
<td>186.1</td>
<td>186.2</td>
<td>186.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Conditional</td>
<td>y_1</td>
<td>46.0 (2.1)</td>
<td>112.0 (3.5)</td>
<td>96.0 (2.7)</td>
<td>186.1 (2.4)</td>
<td>186.2 (2.4)</td>
<td>186.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>y_2</td>
<td>46.0 (0.9)</td>
<td>–</td>
<td>96.2 (3.3)</td>
<td>186.0 (3.8)</td>
<td>184.0 (2.1)</td>
<td>134.1 (2.8)</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>Joint</td>
<td>y_1</td>
<td>45.9 (2.9)</td>
<td>111.7 (3.5)</td>
<td>96.0 (2.9)</td>
<td>186.0 (2.8)</td>
<td>134.0 (2.3)</td>
<td>184.0 (2.3)</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>y_2</td>
<td>46.0 (1.6)</td>
<td>–</td>
<td>95.9 (3.3)</td>
<td>185.9 (3.6)</td>
<td>134.0 (2.6)</td>
<td>184.0 (2.6)</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>Separate</td>
<td>y_1</td>
<td>45.9 (2.9)</td>
<td>112.0 (3.5)</td>
<td>95.9 (2.8)</td>
<td>186.1 (2.7)</td>
<td>134.2 (2.4)</td>
<td>184.0 (2.4)</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>y_2</td>
<td>46.0 (1.6)</td>
<td>–</td>
<td>95.9 (3.2)</td>
<td>185.8 (3.3)</td>
<td>134.0 (2.6)</td>
<td>184.2 (2.4)</td>
<td>64</td>
</tr>
<tr>
<td>300</td>
<td>True position</td>
<td>46</td>
<td>112</td>
<td>96</td>
<td>186</td>
<td>186.1</td>
<td>186.2</td>
<td>186.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Conditional</td>
<td>y_1</td>
<td>46.1 (2.3)</td>
<td>111.9 (3.2)</td>
<td>96.0 (1.7)</td>
<td>185.9 (1.5)</td>
<td>134.0 (2.0)</td>
<td>184.0 (2.0)</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>y_2</td>
<td>46.0 (0.2)</td>
<td>–</td>
<td>96.1 (2.4)</td>
<td>185.8 (3.3)</td>
<td>133.9 (1.1)</td>
<td>184.0 (3.0)</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>Joint</td>
<td>y_1</td>
<td>46.1 (2.4)</td>
<td>112.1 (3.7)</td>
<td>96.2 (2.2)</td>
<td>186.1 (2.1)</td>
<td>134.0 (3.1)</td>
<td>184.0 (1.6)</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>y_2</td>
<td>46.0 (0.7)</td>
<td>–</td>
<td>95.9 (2.8)</td>
<td>185.7 (3.3)</td>
<td>134.0 (1.8)</td>
<td>639.2 (2.0)</td>
<td>180.1 (2.2)</td>
</tr>
<tr>
<td></td>
<td>Separate</td>
<td>y_1</td>
<td>46.1 (2.4)</td>
<td>112.1 (3.7)</td>
<td>96.0 (2.1)</td>
<td>186.2 (2.2)</td>
<td>133.9 (3.2)</td>
<td>134.1 (1.8)</td>
<td>639.2 (1.9)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>y_2</td>
<td>46.0 (0.8)</td>
<td>–</td>
<td>96.0 (2.5)</td>
<td>185.8 (3.3)</td>
<td>134.1 (1.8)</td>
<td>639.2 (1.9)</td>
<td>180.2 (2.1)</td>
</tr>
</tbody>
</table>

Large-scale genetic association analyses

817
Table 2: Mean estimates and standard deviations (in parentheses) of QTL effects obtained with three mapping methods for the simulated data sets with correlation 0.6

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Method</th>
<th>Trait</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q5</th>
<th>Q6</th>
<th>Q7</th>
<th>Q8</th>
<th>Q9</th>
<th>Q10</th>
</tr>
</thead>
<tbody>
<tr>
<td>150</td>
<td>Conditional</td>
<td>y1</td>
<td>0.60</td>
<td>0.21</td>
<td>0.68</td>
<td>0.70</td>
<td>-0.37</td>
<td>1.00</td>
<td>0.00</td>
<td>0.44</td>
<td>0.67</td>
<td>-0.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>y2</td>
<td>-1.40</td>
<td>0.00</td>
<td>0.42</td>
<td>0.24</td>
<td>0.66</td>
<td>-0.70</td>
<td>-0.58</td>
<td>0.65</td>
<td>0.00</td>
<td>0.32</td>
</tr>
<tr>
<td>300</td>
<td>Conditional</td>
<td>y1</td>
<td>0.59 (0.27)</td>
<td>0.27 (0.24)</td>
<td>0.53 (0.19)</td>
<td>0.55 (0.21)</td>
<td>-0.41 (0.22)</td>
<td>0.90 (0.27)</td>
<td>-0.41 (0.39)</td>
<td>0.56 (0.22)</td>
<td>-0.38 (0.23)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>y2</td>
<td>-1.25 (0.27)</td>
<td>-0.36 (0.23)</td>
<td>0.28 (0.23)</td>
<td>0.60 (0.25)</td>
<td>-0.65 (0.23)</td>
<td>-0.50 (0.18)</td>
<td>0.52 (0.18)</td>
<td>0.37 (0.21)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Joint</td>
<td>y1</td>
<td>0.52 (0.33)</td>
<td>0.29 (0.33)</td>
<td>0.57 (0.29)</td>
<td>0.57 (0.30)</td>
<td>-0.38 (0.30)</td>
<td>0.75 (0.31)</td>
<td>0.53 (0.29)</td>
<td>0.56 (0.30)</td>
<td>-0.30 (0.38)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>y2</td>
<td>-1.06 (0.35)</td>
<td>-0.43 (0.30)</td>
<td>0.30 (0.38)</td>
<td>0.59 (0.25)</td>
<td>-0.59 (0.31)</td>
<td>-0.53 (0.29)</td>
<td>0.60 (0.27)</td>
<td>0.38 (0.30)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Separate</td>
<td>y1</td>
<td>0.58 (0.32)</td>
<td>0.32 (0.35)</td>
<td>0.61 (0.28)</td>
<td>0.64 (0.28)</td>
<td>-0.42 (0.31)</td>
<td>0.80 (0.30)</td>
<td>-0.48 (0.34)</td>
<td>0.46 (0.29)</td>
<td>-0.35 (0.35)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>y2</td>
<td>-1.09 (0.33)</td>
<td>-0.45 (0.29)</td>
<td>0.32 (0.36)</td>
<td>0.61 (0.25)</td>
<td>-0.62 (0.29)</td>
<td>-0.56 (0.28)</td>
<td>0.62 (0.27)</td>
<td>0.39 (0.30)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>300</td>
<td>Conditional</td>
<td>y1</td>
<td>0.58 (0.18)</td>
<td>0.25 (0.21)</td>
<td>0.56 (0.18)</td>
<td>0.56 (0.16)</td>
<td>-0.42 (0.16)</td>
<td>0.98 (0.20)</td>
<td>-0.39 (0.16)</td>
<td>0.59 (0.19)</td>
<td>-0.33 (0.19)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>y2</td>
<td>-1.36 (0.20)</td>
<td>-0.37 (0.14)</td>
<td>0.26 (0.18)</td>
<td>0.62 (0.18)</td>
<td>-0.66 (0.19)</td>
<td>-0.51 (0.15)</td>
<td>0.52 (0.16)</td>
<td>0.36 (0.14)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Joint</td>
<td>y1</td>
<td>0.52 (0.23)</td>
<td>0.23 (0.35)</td>
<td>0.55 (0.20)</td>
<td>0.56 (0.25)</td>
<td>-0.36 (0.31)</td>
<td>0.77 (0.29)</td>
<td>-0.41 (0.28)</td>
<td>0.53 (0.31)</td>
<td>-0.29 (0.31)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>y2</td>
<td>-1.17 (0.29)</td>
<td>-0.38 (0.25)</td>
<td>0.29 (0.30)</td>
<td>0.54 (0.22)</td>
<td>-0.56 (0.25)</td>
<td>-0.50 (0.23)</td>
<td>0.52 (0.28)</td>
<td>0.34 (0.26)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Separate</td>
<td>y1</td>
<td>0.55 (0.25)</td>
<td>0.26 (0.39)</td>
<td>0.59 (0.32)</td>
<td>0.58 (0.27)</td>
<td>-0.40 (0.33)</td>
<td>0.80 (0.29)</td>
<td>-0.45 (0.31)</td>
<td>0.57 (0.33)</td>
<td>-0.32 (0.36)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>y2</td>
<td>-1.18 (0.28)</td>
<td>-0.42 (0.27)</td>
<td>0.32 (0.33)</td>
<td>0.57 (0.22)</td>
<td>-0.60 (0.26)</td>
<td>-0.53 (0.22)</td>
<td>0.55 (0.28)</td>
<td>0.34 (0.30)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

QTLs and residual variance for trait of interest. Heritability of the simulated QTLs ranges from 0 to 0.24 for the first trait, whereas from 0 to 0.39 for the second trait. To evaluate the impact of sample size and correlation between the two traits on mapping results, sample size is designated as 150 and 300, and correlation as 0, 0.3, 0.6 and 0.9.

The simulated data sets are analyzed by using conditional mapping method (Conditional for short), joint mapping method (Joint for short) and separate mapping method (Separate for short), respectively. For convenience to compare the three mapping methods, all test statistics are transformed to \( -\log_{10}(p) \), where \( p \) is the probability when the test statistic is greater than the realized value of the test statistic. The simulations are repeated 500 times for estimating QTL parameters and accessing the statistical power of QTL detection. At 5% significance level, statistical power of QTL detection is calculated at each locus as the percentage of the number of those simulations that statistic value exceeds critical value of 1.301 at the locus. Also, false-positive rate is evaluated with the 500 replicated simulations under the null model with zero genetic effects.

The results from three mapping methods are presented in Table 1 for QTL position, Table 2 for QTL genetic effect and Table 3 for statistical power to detect QTL and false-positive rate with correlation \( p = 0.6 \) between the two traits. The results under other correlations are provided in supplementary file. Each mapping method shows general statistical performance: (1) statistical power of QTL detection and the precision of parameter estimation increase as the QTL heritability level increases and (2) large sample size is beneficial to identify QTLs. Under different correlations between the two traits, conditional mapping method is almost identical to joint one in terms of statistical power and QTL parameter estimation. In general, false-positive rates are <10% for all mapping conditions and methods, and the rank in false-positive rate is in the order: conditional mapping method < separate mapping method < joint mapping method, under the same mapping condition. The conditional and joint mapping methods, even without correlation between the two traits, have consistently higher statistical power compared to separate mapping method. This is associated with utilization of the information only on a single trait and not considering the correlation between the two traits by separate mapping method. The differences in statistical power between multivariate and univariate analyses display a trend of enlargement with increasing correlation between the two traits. The estimates for QTL genetic effects are theoretically the same, but practically indistinguishable due to different statistical powers among the three mapping methods.

Actually, the advantage of our proposed method lies in difference in computing run-time increases as well. The computing time for the three mapping methods are recorded for each simulation run (Results not shown). The results showed that, on average, joint mapping method consumes about 10 times more computing time than conditional mapping method under a sample size of 300. The difference in computing time-consuming gets larger as the number of traits increases.
Real data analysis

For mapping eQTL in yeast, 112 meiotic recombinant progenies were generated from a laboratory strain and a natural isolate strain [41]. On each individual, expression profiles of 6216 gene expression traits were measured and 2956 SNP markers were genotyped. The description of the data set can be found at Brem and Kruglyak [41]. The data set is re-analyzed by using our method to understand the genetic architecture of gene expressions. The results are compared with those obtained with separate mapping method. However, joint mapping method is not able to simultaneously analyze those gene expressions.

The total number of detectable QTLs is basically the same between the two mapping methods, 879 of which are simultaneously identified by using the two mapping methods (83% overlapping). The distribution of QTL numbers in gene expressions is listed in Table 4s. Total gene expressions with more than six QTLs obtained by conditional mapping method are greater than those by separate mapping method. The number of gene expressions without QTL is 4694 for conditional mapping method, whereas 4712 for separate mapping method. Especially, conditional mapping method finds more pleiotropic QTLs than separate mapping method (617 pleiotropic QTLs for conditional mapping method and 582 for separate mapping method).

Figure 1 shows the distribution in total QTL heritability of 6216 gene expressions. As can be seen, most of gene expressions have total QTL heritability of 0.6 for the two mapping methods. However, conditional mapping method identifies more gene expressions with total QTL heritability over 0.6 than separate mapping method. For all analyzed gene expressions, conditional mapping method detects 28 QTLs with heritability of >0.9, of which, the largest QTL for gene expression ‘YDL227C’ has 0.978 heritability on chromosome 4. The 98.7% phenotypic variation of gene expression ‘YGL016W’ is explained by the two SNPs, which are SNP ‘NBR034W’ and SNP ‘YBR161W’ on chromosome 2. The largest QTLs are also obtained with separate mapping method.

As an illustration, correlation architecture and genetic control for gene expression are displayed in Figure 2. By the LASSO in step (2), gene expression ‘YDL154W’ is statistically associated with other 12 gene expressions including YNL250W, YDR498C,

Table 3: Statistical powers of QTL detection and false positive rates (FPR) obtained with three mapping methods for the simulated data sets with correlation 0.6

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Method</th>
<th>Trait</th>
<th>Statistical power</th>
<th>FPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>150</td>
<td>Conditional</td>
<td>y1</td>
<td>890 44.2 84.0 870 72.4 98.8 0.0 64.2 90.4 56.8 4.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>y2</td>
<td>998 0.0 60.0 40.0 88.6 92.2 80.0 82.4 0.0 55.0 4.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Joint</td>
<td>y1</td>
<td>870 46.2 82.4 88.8 72.0 97.6 0.0 63.6 89.2 57.0 6.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>y2</td>
<td>95.8 0.0 59.8 40.4 89.8 91.0 79.4 83.6 0.0 54.8 6.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Separate</td>
<td>y1</td>
<td>60.6 33.8 59.6 65.4 39.4 81.0 0.0 49.8 65.4 40.2 5.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>y2</td>
<td>89.6 0.0 45.4 37.4 63.4 62.6 56.2 63.6 0.0 39.6 4.8</td>
<td></td>
</tr>
<tr>
<td>300</td>
<td>Conditional</td>
<td>y1</td>
<td>97.4 54.4 92.6 96.8 84.8 99.8 0.0 77.0 97.6 70.8 4.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>y2</td>
<td>100.0 0.0 68.6 48.4 97.8 98.6 95.8 95.0 0.0 73.0 3.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Joint</td>
<td>y1</td>
<td>96.8 61.2 92.6 95.8 83.8 99.0 0.0 77.6 98.0 71.0 6.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>y2</td>
<td>99.6 0.0 70.2 52.4 97.4 98.4 94.0 96.0 0.0 72.6 5.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Separate</td>
<td>y1</td>
<td>70.2 41.2 76.2 77.4 51.2 91.8 0.0 57.2 74.0 43.6 5.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>y2</td>
<td>98.6 0.0 49.8 37.0 70.8 75.2 66.4 72.4 0.0 44.0 4.8</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: The distribution of gene expressions controlled by different numbers of QTLs

<table>
<thead>
<tr>
<th>QTL numbers</th>
<th>Conditional</th>
<th>Separate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>606</td>
<td>559</td>
</tr>
<tr>
<td>2</td>
<td>320</td>
<td>310</td>
</tr>
<tr>
<td>3</td>
<td>208</td>
<td>214</td>
</tr>
<tr>
<td>4</td>
<td>150</td>
<td>165</td>
</tr>
<tr>
<td>5</td>
<td>84</td>
<td>118</td>
</tr>
<tr>
<td>6</td>
<td>72</td>
<td>63</td>
</tr>
<tr>
<td>7</td>
<td>30</td>
<td>34</td>
</tr>
<tr>
<td>8</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>9</td>
<td>20</td>
<td>12</td>
</tr>
<tr>
<td>&gt;10</td>
<td>17</td>
<td>13</td>
</tr>
<tr>
<td>Total</td>
<td>1522</td>
<td>1504</td>
</tr>
</tbody>
</table>
YDR416W, YJL160C, YPR110C, YMR133W, YKR057W, YJR056C, YPL274W, YLR174W, YFL018C and YDL180W. It is genetically determined by five SNPs, which are YDR144C on chromosome 4, YFR003C on chromosome 6, YNL318C and YNL158W on chromosome 14 and NOL009W on chromosome 15 as well. The heritability of the SNPs ranges from 0.08 to 0.16.

**DISCUSSION**

Large-scale genotype–phenotype association analyses are realized by transforming joint association analysis for the correlated phenotypes to separate association analyses for conditional phenotypes. For implementation of this method, the LASSO with coordinate descent step [39, 40] is applied to efficiently estimate conditional phenotype and sparse non-zero genetic effects of each trait. Then, ordinary least square method is used to unbiasedly estimate the QTL genetic effects and to statistically infer the significance of QTLs for each trait. In estimating conditional phenotype of each trait, additionally, the correlation architecture between each trait and other traits is provided by shrinkage estimation of each conditional phenotype.

Our proposed method is essentially equivalent to the staged solution for structural equation model, in which dependent variables are phenotypes of each trait, and independent variables are genetic effects and the residuals of other traits. In usual QTL

---

**Figure 1:** The distribution in total QTL heritability of 6216 gene expressions in yeast. In the plot, solid bar is for conditional mapping method and hollow bar for separate mapping method.

**Figure 2:** Correlation architecture and genetic control for gene expression ‘YDL154W’ in yeast.
mapping, however, the phenotypes of other traits are used as independent variables of the structural equation model [35, 42]. As the independent residual in Equation (6) is derived from the regression of the residual of each trait on the residuals of other traits in Equation (1), it insures that the estimates of genetic effects obtained with our proposed method are comparable with those with joint association analysis from the viewpoint of structural equation model. Simulation indicates that conditional mapping method proposed here performs as well as the joint one in terms of both parameter estimation and power to detect QTL.

Conditional genetic analysis has long been used for estimating genetic variances. Henderson [43, 44] estimated conditional genetic variance given selected base population. Zhu [45] analyzed conditional genetic effects and variance components in developmental genetics. No attempt is made to locate QTLs for multiple traits. With thousands of SNPs, the knowledge about genes and pathways can also be incorporated into GWAS [46, 47] for each conditional phenotype by using grouped LASSO [39, 40]. Of course, our method is also appropriate for analyzing subgroups of clustered phenotypes. Such an approach that considers the information about all genotypes and phenotypes will potentially increase the power of detecting weak associations and reduce susceptibility to noise [36].

SUPPLEMENTARY DATA
Supplementary data are available online at http://bib.oxfordjournals.org/.

Key points
- By defining conditional phenotype of each trait given other traits, large-scale genotype–phenotype association analyses are transformed to separate genotype-conditional phenotype ones.
- For implementation of this method, the LASSO with coordinate descent step is applied to efficiently estimate conditional phenotype and sparse non-zero genetic effects of each trait.
- Correlation architecture among traits can also be provided by shrinkage estimation, given each conditional phenotype.
- The proposed method is evaluated with extensive computer simulations and real data analysis for mapping eQTLs in yeast.

FUNDING
National Natural Science Foundation of China (in part) (30972077 and 31172190 to R.Y.).

References


