A Stochastic Expectation and Maximization (SEM) Algorithm for Detecting Quantitative Trait Associated Genes

Haimao Zhan¹, Xin Chen² and Shizhong Xu¹,*

¹ Department of Botany and Plant Sciences, University of California, Riverside, CA 92521 USA
² Department of Statistics, University of California, Riverside, CA 92521, USA

ABSTRACT

Motivation: Most biological traits may be correlated with the underlying gene expression patterns that are partially determined by DNA sequence variation. The correlations between gene expressions and quantitative traits are essential for understanding the functions of genes and dissecting gene regulatory networks.

Results: In the present study, we adopted a novel statistical method, called the stochastic expectation and maximization (SEM) algorithm, to analyze the associations between gene expression levels and quantitative trait values and identify genetic loci controlling the gene expression variations. In the first step, gene expression levels measured from microarray experiments were assigned to two different clusters based on the strengths of their association with the phenotypes of a quantitative trait under investigation. In the second step, genes associated with the trait were mapped to genetic loci of the genome. Because gene expressions are quantitative, the genetic loci controlling the expression traits are called expression quantitative trait loci (eQTL). We applied the same SEM algorithm to a real dataset collected from a barley genetic experiment with both quantitative traits and gene expression traits. For the first time, we identified genes associated with eight agronomy traits of barley. These genes were then mapped to seven chromosomes of the barley genome. The SEM algorithm and the result of the barley data analysis are useful to scientists in the areas of bioinformatics and plant breeding.

Availability and implementation: The R program for the SEM algorithm can be downloaded from our website: http://www.statgen.ucr.edu
Contact: shizhong.xu@ucr.edu
Supplementary information: Supplementary data are available in the Bioinformatics online system.

1 INTRODUCTION

Differential expression analysis often applies to discrete phenotypes (primarily dichotomous phenotypes). The phenotype is often defined as "normal" or "affected". If a phenotype is measured quantitatively, it is often converted into two or a few discrete or-
problems have been encountered for this method. One is the identifiability problem where the cluster labels can be exchanged among different clusters. The other problem occurs when two or more clusters often have the same cluster mean. These two problems can be avoided by introducing a small noise to the cluster means in every iterative step of the EM algorithm (Qu and Xu, 2006). This ad hoc modification of the EM algorithm lacks strong theoretical foundation and cannot guarantee to produce the optimal result. In this study, we proposed an alternative method with a rigorous theoretical basis to solve the problem. We used this new method to detect expressed genes that are associated with multiple quantitative traits of barley.

The gene expression levels are quantitative traits (Morley, et al., 2004). Finding the genetic loci controlling the expressions can help identify gene regulation networks (Cookson, et al., 2009). Trait specific gene networks can be inferred by studying the genetic loci controlling the expressed genes only associated with the trait under investigation. In this study, we focus on a new method called the stochastic expectation and maximization (SEM) algorithm (Celeux and Diebolt, 1986) and its application to both eQTL mapping and phenotype associated microarray data analysis. This new method is then compared with the existing EM algorithm (Jia and Xu, 2005) to demonstrate its superiority. A real dataset collected in the North American Barley Genome Project (NABGP) is used for the demonstration.

2 MATERIALS AND METHODS

2.1 Experimental data

The gene expression data were published by Luo et al. (2007) and downloadable from the ArrayExpress: http://www.ebi.ac.uk/microarray-as/aer/entry with accession number: E-TABM-112. The phenotypic values of eight quantitative traits of barley were published by Hayes et al. (1993) and downloadable from the following website: http://wheat.pli.usda.gov/ggpages/SxM/phenotypes.html. Detailed description of the experiment can be found from the original study (Hayes, et al., 1993). The experiment involved 150 double haploid (DH) lines derived from the cross of two spring barley varieties, Morex and Steptoe. All the 150 DH lines were microarrayed for 22840 transcripts. The eight traits are α-amylase, diastatic power, grain protein, grain yield, heading date, plant height, lodging and malt extract. The phenotypes of the traits were measured in different environments (locations and years). The number of replicated measurements ranged from 6 to 16 depending on different traits. Both the single trait association and multiple trait joint association analyses were conducted for all the eight traits using the average trait values across all environments.

The original (raw) microarray data were normalized using the RMA algorithm (Irizarry, et al., 2003) implemented in the GeneSpring GX 11 software package (Agilent Technologies, Santa Clara, CA). ArrayExpress also provides the preprocessed dataset without log transformation. The phenotypic values of each trait were rescaled so that the range of each trait is between -1 and +1. The formula for the rescaling is

\[ Z_k = \frac{X_k - \bar{X}_{min} - \bar{X}_{max}}{X_{max} - X_{min}} - 1 \]

where \( X_k \) is the original phenotypic value for the \( k \)th line, \( X_{min} \) and \( X_{max} \) are the minimum and maximum values of the phenotypic value, respectively, and \( Z_k \) is the rescaled phenotypic value for \( k = 1, \ldots, n \), where \( n \) is the sample size (number of DH lines).

2.2 Linear Model

Denote the microarray data by a data matrix \( Y \) with \( n \) rows and \( m \) columns, where \( n \) is the number of individuals subject to the microarray analysis and \( m \) is the number of microarrayed genes. Let \( y_{ijk} \) be the \( j \)th column of matrix \( Y \), i.e., an \( n \times 1 \) vector for the expression levels of gene \( j \) for all the \( n \) subjects ( \( j = 1, \ldots, m \) ). Let \( Z \) be an \( n \times q \) matrix for the rescaled phenotypic values of \( q \) quantitative traits measured from all \( n \) individuals. Let \( X \) be an \( n \times p \) matrix for some factors not directly relevant to the quantitative traits, for example, gender effect, age effect and so on. We now have three sources of data, (1) \( Y \) the microarray data, (2) \( Z \) the phenotypic data, and (3) \( X \) the cofactors not directly relevant to the association study. The cofactors are not something of interest themselves, but may affect the gene expressions. They are included in the model to reduce or eliminate the interference on the association between \( Y \) and \( Z \). The expressed levels of gene \( j \) can be expressed using the following linear model,

\[ y_{ijk} = x_j \beta_j + z_{ijk} + \epsilon_{ijk} \]

where \( \beta_j \) is a \( p \times 1 \) vector for the effects of cofactors, \( z_{ijk} \) is a \( q \times 1 \) vector for the regression coefficients of gene \( j \) on all the \( q \) traits. The residual error \( \epsilon_{ijk} \) is an \( n \times 1 \) vector with an assumed \( N(0, \sigma^2) \) distribution. This assumption is very common in the linear model analysis. The DH plants are indeed independent samples from the line cross of the barley experiment. In the special case of one phenotype with no cofactors, \( p = q = 1 \) and \( X \) is a vector of unity with a dimensionality of \( n \times 1 \).

We now assign prior distributions to the parameters included in the linear model. The prior distribution for \( \beta_j \) is

\[ \beta_j \sim N(\mu_j, \Sigma_j) \]

where \( \mu_j \) is a \( p \times 1 \) vector of mean and \( \Sigma_j \) is an unknown \( p \times p \) positive definite variance matrix. The prior distribution of \( z_{ijk} \) is a Gaussian mixture with two components,

\[ z_{ijk} \sim N(0, \Sigma) + (1 - \pi) N(0, \Sigma_j) \]

In the above Gaussian mixture, \( \Sigma = \rho I \) is a known diagonal matrix with a common \( \rho = 10^{-4} \) across all the diagonal elements. In other words, \( \Sigma_j \) is a known positive definite matrix with values close to zero and the value can be changed according to the investigator’s preference. For the other cluster, \( \Sigma_j \) is an unknown positive definite variance matrix. This Gaussian mixture prior divides all the genes into two clusters, one (cluster 1) being associated with the phenotypes and the other (cluster 0) not associated with the phenotype. Variable \( \pi \) is a prior probability that a gene randomly selected from the pool belongs to cluster 1. A gene classified into cluster 1 is claimed to be associated with the traits while genes classified into cluster zero are not associated with the traits. The actual parameters involved in the problem are denoted by

\[ \theta = \{ \mu_j, \Sigma_j, \pi \} \]

We are also interested in \( \rho \), the posterior probability of gene \( j \) being associated with the traits. The relationship between the posterior \( \rho \) and the prior \( \pi \) will be presented later.

Let \( \delta_j \sim \text{Bernoulli}(\rho_j) \), for \( 0 < \rho_j < 1 \), be an indicator variable for the cluster membership of gene \( j \). It is defined as

\[ \delta_j = \begin{cases} 1 & \text{if } j \text{ belongs to cluster 1} \\ 0 & \text{if } j \text{ belongs to cluster 0} \end{cases} \]

Given \( \delta_j \), the genetic effect \( \gamma_j \) has the following distribution,

\[ \gamma_j \sim \delta_j N(0, \Sigma) + (1 - \delta_j) N(0, \Sigma_j) \]
Note that this conditional distribution is not a Gaussian mixture because the membership is already known. The indicator variable \( \delta_j \) tells which Gaussian component \( y_j \) belongs to. Also note that the parameter vector does not include \( \beta_j \) and \( y_j \). Under the random model framework, \( \beta_j \) and \( y_j \) are treated as missing values. If they are integrated out, the density of \( y_j \) given the cluster membership is normal with mean and variance shown in the following normal density

\[
p(y_j | \theta, \delta) = N(y_j | X \mu_j, X \Sigma_j X^T + Z \Theta_j Z^T + I \sigma^2)
\]

where

\[
\Theta_j = \delta_j \Sigma_j + (1-\delta_j) \Sigma_0
\]

and the notation \( N[y_j | a, b] \) stands for the normal density of variable \( y_j \) with mean \( a \) and variance \( b \). Given the cluster membership, \( \delta = [\delta_j] \), the log likelihood function for the entire dataset is

\[
L(\theta | \delta) = \sum_{m=1}^m \ln [p(y_j | \theta, \delta_j)]
\]

The MLE of parameters are obtained through a two-step approach. The first step is to estimate the parameters by maximizing the above log likelihood function given \( \delta = [\delta_j] \) through the regular expectation-maximization (EM) algorithm (Dempster et al., 1977). This step is called the EM step. The second step is to stochastically simulate \( \delta = [\delta_j] \) from its conditional posterior distribution. This step is called the stochastic step. The two steps are repeated iteratively until a stationary distribution is reached for each parameter, an algorithm called the stochastic expectation and maximization (SEM) algorithm (Celeux and Diebolt, 1986). The stochastic step and the EM step are performed sequentially, not in parallel.

### 2.3 Stochastic sampling

The density of \( y_j \) defined in equation (8) can be split into the following two densities,

\[
p_j(y_j | \theta) = p(y_j | \theta, \delta_j = 1) = N(y_j | X \mu_j, X \Sigma_j X^T + Z \Theta_j Z^T + I \sigma^2)
\]

and

\[
p_0(y_j | \theta) = p(y_j | \theta, \delta_j = 0) = N(y_j | X \mu_j, X \Sigma_j X^T + Z \Theta_j Z^T + I \sigma^2)
\]

The posterior probability that \( \delta_j = 1 \) is

\[
\rho_j = E(\delta_j | \theta, y_j) = \frac{\pi p_j(y_j | \theta)}{\pi p_j(y_j | \theta) + (1-\pi)p_0(y_j | \theta)}
\]

Because \( \delta_j \) is a Bernoulli variable, it is sampled from

\[
p(\delta_j) = \text{Bernoulli}(\delta_j | \rho_j)
\]

distribution. Once \( \delta = [\delta_j] \) are sampled for all genes in the stochastic process, we can proceed with the EM algorithm described below.

### 2.4 EM algorithm

The EM algorithm for the Gaussian mixture model is standard (Dempster et al., 1977) and thus we only provide the EM steps without proof. Denote the variance covariance matrix of \( \gamma_j \) conditional on \( \delta_j \) by

\[
\Theta_j = \text{var}(y_j | \delta_j) = \delta_j \Sigma_j + (1-\delta_j) \Sigma_0
\]

Let us define

\[
V_j = X \Sigma_j X^T + Z \Theta_j Z^T + I \sigma^2
\]

We now provide the formulas for updating each parameter using the EM algorithm. Given \( \delta_j \), the updated proportion of genes coming from cluster 1 is

\[
\pi = \frac{1}{m} \sum_{j=1}^m \delta_j
\]

The population mean of \( \mu_j \) is updated using

\[
\mu_j = \frac{\sum_{j=1}^m X^T V_j^{-1} y_j}{\sum_{j=1}^m X^T V_j^{-1} x_j}
\]

The variance-covariance matrix of \( \beta_j \) is denoted by \( \Sigma_j \) and updated using

\[
\Sigma_j = \frac{1}{m} \sum_{j=1}^m E(\beta_j, \beta'_j) = \frac{1}{m} \sum_{j=1}^m \left[ E(\beta_j) E(\beta'_j) + \text{var}(\beta_j) \right]
\]

where

\[
E(\beta_j) = \Sigma_j Z^T V_j^{-1} (y_j - X \mu_j)
\]

and

\[
\text{var}(\beta_j) = \Sigma_j - \Sigma_j Z^T V_j^{-1} \Sigma_j
\]

Given \( \delta_j \), the unknown variance-covariance matrix of \( \gamma_j \) is

\[
\Theta_j = \delta_j \Sigma_j + (1-\delta_j) \Sigma_0
\]

However, the corresponding matrix \( \Sigma_0 \) is a constant. Therefore, we only need to update \( \Sigma_j \) using all \( y_j \) that come from cluster one. The updated equation for \( \Sigma_j \) is

\[
\Sigma_j = \frac{1}{\pi m} \sum_{j=1}^m \delta_j E(\gamma_j, \gamma'_j) = \frac{1}{\pi m} \sum_{j=1}^m \left[ E(\gamma_j) E(\gamma'_j) + \text{var}(\gamma_j) \right]
\]

where

\[
E(\gamma_j) = \Sigma_j Z^T V_j^{-1} (y_j - X \mu_j)
\]

and

\[
\text{var}(\gamma_j) = \Sigma_j - \Sigma_j Z^T V_j^{-1} \Sigma_j
\]

### Table 1: Numbers of genes associated with individual traits in the barley microarray data analysis using the two-cluster SEM algorithm and the three-cluster EM algorithm.

<table>
<thead>
<tr>
<th>Trait</th>
<th>SEM Algorithm</th>
<th>EM Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number</td>
<td>Proportion</td>
</tr>
<tr>
<td>Alpha-amylose</td>
<td>644</td>
<td>2.82%</td>
</tr>
<tr>
<td>Diastatic power</td>
<td>467</td>
<td>2.04%</td>
</tr>
<tr>
<td>Grain protein</td>
<td>888</td>
<td>3.89%</td>
</tr>
<tr>
<td>Grain yield</td>
<td>457</td>
<td>2.04%</td>
</tr>
<tr>
<td>Heading date</td>
<td>401</td>
<td>1.76%</td>
</tr>
<tr>
<td>Height</td>
<td>784</td>
<td>3.43%</td>
</tr>
<tr>
<td>Lodging</td>
<td>650</td>
<td>2.85%</td>
</tr>
<tr>
<td>Malt extract</td>
<td>526</td>
<td>2.30%</td>
</tr>
</tbody>
</table>
The residual error variance is updated using

\[ \sigma^2 = \frac{1}{mn} \sum_{j=1}^{m} \left( y_j - X \mu_j \right)^T \left( y_j - X \mu_j - X E(\beta_j) - \delta, ZE(\gamma_j) \right) \]  

(25)

The E-step of the EM algorithm consists of calculating \( E(\beta_j), E(\gamma_j), \) var(\( \beta_j \)) and var(\( \gamma_j \)). The M-step consists of calculating \( \theta = \{ \mu^0, \Sigma^0, \Sigma^0, \sigma^2, \pi \} \). So far all parameter have been updated. We can now combine the stochastic steps with the EM steps to compute the analysis. Note again that the stochastic and EM steps are performed sequentially and repeated many times until a stationary distribution for each parameter is reached.

### 2.5 SEM estimate

The SEM algorithm differs from the classical EM algorithm in that the parameters do not converge to some fixed values; rather, they converge to a stationary distribution due to the stochastic process of \( \delta = [\delta_i] \). We can monitor the converging process for each parameter. Once all parameters have converged, we start to collect the posterior sample for \( \theta \). Unlike the posterior sample for the fully Bayesian analysis, the observations with the posterior sample for the SEM algorithm are not correlated. The posterior sample size, denoted by \( T \), does not have to be large; \( T = 100 \) seems to be sufficient. Let \( \theta^{(t)} = \{ \mu^{(t)}, \Sigma^{(t)}, \sigma^{(t)}, \pi^{(t)} \} \) be the \( t \)-th observation in the posterior sample (after convergence), the estimate parameter vector of the SEM algorithm is

\[ \hat{\theta} = \frac{1}{T} \sum_{i=1}^{T} \theta^{(i)} \]  

(26)

The most important quantity of the SEM analysis is not the entire vector of \( \theta \); rather, it is

\[ \hat{\rho}_j = \frac{1}{T} \sum_{i=1}^{T} \rho^{(i)}_j = \frac{1}{T} \sum_{i=1}^{T} \rho^{(i)}_j \]  

(27)

that is most important because it represents the posterior probability that gene \( j \) belongs to cluster 1, i.e., gene \( j \) is associated with the quantitative traits. These \( \hat{\rho}_j \)'s are ranked in a descending order. The top proportion of genes is selected as candidate genes associated with the phenotypes. The proportion is defined by the investigator in an arbitrary manner. Some objective criteria, e.g., false discovery rate (FDR), may be used, but it is not the focus of this study. Here, we simply chose an arbitrary cut-off value of \( \rho_0 \geq 0.9 \) to declare significant association.

### 2.6 Expression quantitative trait locus (eQTL) mapping

The gene expression levels can be treated as quantitative traits and QTL mapping can be performed on each transcript, so called eQTL mapping (Kendziorski et al., 2006). The problem with the eQTL analysis is that the large number of expression traits make eQTL mapping very difficult. The SEM algorithm developed for the quantitative trait associated microarray data analysis can be extended to eQTL mapping with limited modification. This section describes the application of the SEM algorithm to eQTL mapping.

Consider \( Q \) markers with known map positions and the genotypes for all the \( n \) individuals. We will study the association of all the \( m \) transcripts simultaneously with the \( k \)-th marker for \( k = 1, \ldots, Q \). The approach is similar to the interval mapping in which one marker is studied at a time (Lander and Botstein, 1989). The entire eQTL mapping will take \( Q \) separate analyses. Using the same model as given in equation (2), we now replace the phenotype \( Z \) by the numerically coded genotype of marker \( k \) denoted by \( Z_k \) so that

\[ y_i = X \beta + Z_k \gamma + \epsilon_i \]  

(28)

where \( \gamma_k \) is the QTL effect for transcript \( j \) at marker \( k \). The \( Z_k \) variable is defined as

\[ Z_k = \begin{cases} +1 & \text{for } A_A, \\ -1 & \text{for } A_A \end{cases} \]  

(29)

In addition, the posterior probability of gene \( j \) associated with marker \( k \) is denoted by \( \rho_{jk} \). These parameters \( \{ \pi, \rho \} \) are important to the eQTL mapping. The SEM algorithm remains the same as before except that we must analyze the data \( Q \) times (one for each marker).

### 3 RESULTS

#### 3.1 Single trait association

Since each trait was measured from multiple environments, we took the average of the phenotypic values across the environments as the phenotypic values that entered the linear model for analysis. Therefore, the \( Z \) matrix for each trait analysis was an \( n \times 1 = 150 \times 1 \) vector for the average phenotypic values (rescaled between -1 and +1). The number of transcripts (genes) measured in the experiment was \( m = 22840 \). Since no other cofactor existed except the intercept, \( \beta_j \) is a scalar with dimension \( p = 1 \). For the single trait analysis, \( q = 1 \) for each trait analysis. Each of the five parameters \( \theta = \{ \mu_0, \Sigma_0, \Sigma_1, \sigma^2, \pi \} \) is of single dimension.

Both the (two-cluster) SEM algorithm developed here and the three-cluster EM algorithm of Jia and Xu (2005) were used for the single trait association study. The EM algorithm of Jia and Xu (2005) classified each gene into one of three clusters. The three
clusters shared a common variance of the regression coefficient but with three different means. The three cluster means were restricted with negative value for the first cluster, zero for the second cluster and positive value for the third cluster. Genes classified into either the first or the third cluster are differentially expressed genes. The criterion of detection for each gene was that the posterior probability of being in cluster 2 (the neutral cluster) was less than 0.1. This is equivalent to the criterion of \( \rho_j \geq 0.9 \) in the SEM analysis.

The numbers of genes associated with each of the eight traits are listed in Table 1 for the SEM and the EM algorithms. We can see that the SEM algorithm consistently detected more genes than the EM algorithm. The result of the SEM algorithm shows that more genes are associated with the height and grain protein than other traits. The heading date trait has the least number of associated genes. The estimated parameters for all the eight traits obtained with the SEM algorithm are given in Table S2 (Sheet1-Sheet8) for interested readers.

In order to compare the performance of the SEM model and the EM model, we carried out a simulation experiment based on the barley data (grain yield) analysis. The ROC curve (Figure 1) shows that both the SEM and the EM models have high sensitivity and specificity, but the SEM model performs better and has higher sensitivity. It is well known that the EM algorithm tends to converge to some sub-optimal values which are close to the initial values. We set different parameters and simulated several datasets based on the SEM model to test the convergence of parameters. When parameters are precisely estimated, the EM model is able to identify most of the associated genes. If parameters converge to some local optimal values that are different from the true values, the sensitivity is quite low. However, the SEM model can identify most associated genes in both cases and has high sensitivity and specificity, which are similar to the EM results when parameters were estimated well. We also performed permutation for the real data analysis by permuting the phenotypes (grain yield) to test the specificity of the two methods. After 100 permutations by the grain yield, we took average of the posterior probabilities generated from the 100 permutations for each gene and we found that no gene had probabilities exceeding the cut-off point (0.9), which indicates that the SEM method also has good specificity in real data analysis.

We choose the grain yield as an example to demonstrate the converging process of the estimated parameters. The trace plots (parameters against the iteration) are depicted in Figure 2 for all the five parameters and the regression coefficient of gene AF250937_s_at. From the trace plots, we can see that all parameters have converged in about 10 iterations. After the parameters converged to their stationary distributions, the parameters fluctuated around the mean values and the means are the SEM estimates of the parameters.

We also presented the predicted regression coefficients obtained with the SEM analysis and the scatter plots of the observed genes expressions for the genes associated with each trait (Figure S1). Some genes have positive associations with the traits, e.g., Contig23592_at and Contig3295_at. The rest of the parameters, \( \pi \) and \( \sigma^2 \), remain scalars. The Z matrix is an 150×8 matrix for all the eight traits measured from all the 150 DH lines. Only the SEM algorithm was used in this analysis because the EM algorithm of Jia and Xu (2005) cannot be applied to multiple trait association study.

Using the same criterion of \( \rho_j \geq 0.9 \) to declare significance association, we detected a total of 1646 genes that are jointly associated with the eight traits, accounting for 7.2% of all the 22840 genes included in the analysis. The list of associated genes is given in the supplemental data (Table S3) for interested readers. The estimated parameters are \( \hat{\beta} = 0.089 \) and \( \hat{\sigma}^2 = 0.067 \) for the proportion and residual error variance, respectively. The estimated \( \hat{\mu}_p \) and \( \Sigma_p \) are \( \hat{\mu}_p = 7.978 \) and \( \Sigma_p = 4.652 \), respectively. The estimated variance matrix of the differentially expressed cluster is

### Table 2. The estimated regression coefficients for the top ten genes jointly associated with all eight traits. The partial regression coefficients of gene on the eight traits are denoted by \( \hat{\gamma}_i \), \( i = 1, \ldots, 8 \).

<table>
<thead>
<tr>
<th>Rank</th>
<th>Probe set ID</th>
<th>( \gamma_1 )</th>
<th>( \gamma_2 )</th>
<th>( \gamma_3 )</th>
<th>( \gamma_4 )</th>
<th>( \gamma_5 )</th>
<th>( \gamma_6 )</th>
<th>( \gamma_7 )</th>
<th>( \gamma_8 )</th>
<th>F-test</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Contig1132_s_at</td>
<td>-2.99</td>
<td>0.99</td>
<td>3.42</td>
<td>-1.99</td>
<td>-0.25</td>
<td>0.75</td>
<td>-0.40</td>
<td>-0.05</td>
<td>945.744</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>2</td>
<td>Contig2163_at</td>
<td>2.32</td>
<td>-0.68</td>
<td>-2.93</td>
<td>1.76</td>
<td>0.38</td>
<td>-0.55</td>
<td>0.41</td>
<td>0.12</td>
<td>581,689</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>3</td>
<td>Contig1524_at</td>
<td>-0.12</td>
<td>-0.77</td>
<td>-1.88</td>
<td>0.62</td>
<td>-0.89</td>
<td>0.45</td>
<td>2.15</td>
<td>1.62</td>
<td>543,998</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>4</td>
<td>Contig2279_at</td>
<td>0.71</td>
<td>-0.83</td>
<td>1.18</td>
<td>-0.10</td>
<td>2.30</td>
<td>0.61</td>
<td>1.62</td>
<td>-1.63</td>
<td>492,936</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>5</td>
<td>Contig4769_at</td>
<td>0.10</td>
<td>0.35</td>
<td>-0.72</td>
<td>0.80</td>
<td>-0.82</td>
<td>-2.05</td>
<td>-1.51</td>
<td>4.94</td>
<td>481,062</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>6</td>
<td>Contig11126_at</td>
<td>2.67</td>
<td>-0.18</td>
<td>-2.32</td>
<td>1.23</td>
<td>0.55</td>
<td>-0.82</td>
<td>0.28</td>
<td>0.21</td>
<td>435,547</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>7</td>
<td>Contig23772_s_at</td>
<td>1.81</td>
<td>-0.67</td>
<td>-2.53</td>
<td>1.50</td>
<td>0.13</td>
<td>-0.55</td>
<td>0.59</td>
<td>0.003</td>
<td>421,314</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>8</td>
<td>HVSME000020F06r2_at</td>
<td>0.55</td>
<td>-0.74</td>
<td>0.92</td>
<td>-0.16</td>
<td>2.16</td>
<td>0.75</td>
<td>1.33</td>
<td>-1.45</td>
<td>396,722</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>9</td>
<td>Contig11126_at</td>
<td>-0.44</td>
<td>-0.21</td>
<td>0.62</td>
<td>-0.86</td>
<td>0.88</td>
<td>1.91</td>
<td>1.09</td>
<td>-4.10</td>
<td>438,061</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>10</td>
<td>Contig4526_at</td>
<td>-0.45</td>
<td>-1.17</td>
<td>1.11</td>
<td>0.88</td>
<td>1.27</td>
<td>1.38</td>
<td>0.58</td>
<td>-1.35</td>
<td>347,999</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>
Fig. 2. The SEM convergence processes of five parameters for the grain yield trait and the regression coefficient ($γ$) of gene AF250937_s_at (the gene having strong association with grain yield).

Note that the proportion of genes detected (7.2%) in the experiment is not the same as $\hat{π} = 8.9\%$ because the former depends on the cut-off point ($ρ_j \geq 0.9$) used for gene declaration whereas the latter represents the probability that a randomly selected gene belongs to the associated cluster and it does not depend on the cut-off point. In the simulation study, we used parameters estimated from the barley data to simulate 5000 genes (100 individuals) based on the SEM model. The SEM algorithm identified all associated genes, which indicated that SEM does have high sensitivity and specificity. By randomly permuting the eight traits, we tested the specificity of SEM in real data analysis. Among the total of 22840 genes, 521 genes still had significant effects in the permuted data. The false positive is $521/22840 = 0.022811$, reasonably low.

Interestingly, all genes associated with the 8 traits in the single trait analysis were also detected in the joint analysis, demonstrating the high efficiency of the joint analysis. The estimated regression coefficients for the top ten genes jointly associated with all traits are listed in Table 2 along with the F-test statistics and the p-values. The F-test statistic was calculated using

$$F_j = \frac{1}{q} \hat{ρ}_j^2 \left( \hat{Σ}_i - \hat{Σ}_i Z_j V_j Z_j^T \hat{Σ}_i \right)^{-1} \hat{γ}_j$$

(31)

where

$$\hat{γ}_j = \hat{Σ}_i Z_j^T \hat{V}_j (y_j - X \hat{μ}_β)$$

(32)

and

$$\hat{V}_j = X \hat{Σ}_i X^T + Z \hat{Σ}_i Z^T + I \hat{σ}^2$$

(33)

The p-value was calculated using

$$p-value = 1 - F_{\hat{σ}, \hat{σ}}(\hat{F}_j)$$

(34)

where $F_{\hat{σ}, \hat{σ}}(x)$ is the cumulative distribution function of the central F distribution with numerator degree of freedom $q = 8$ and the denominator degree of freedom $\infty$.

3.3 Expression quantitative trait locus (eQTL) mapping

The purpose of the eQTL mapping is to identify the locations of the genome that control the expressions of the transcripts. We used the results of the previous analysis to reduce the number of transcripts for the eQTL analysis. For example, out of the 22840 transcripts, we identified 888 genes that are associated with the grain protein trait. The eQTL mapping was then targeted to these 888 transcripts. This has dramatically reduced the number of transcripts for eQTL mapping related to the grain protein trait. Recall that Table 1 gives the number of transcripts associated with each of the eight traits. The eQTL mapping for each trait was only conducted on the identified transcripts. The barley genome contains seven chromosomes. The total number of SNP markers investigated was $Q = 495$ with an average marker interval less than 2 centiMorgan, covering the entire barley genome.

The entire eQTL analysis took about nine hours of Intel Core Duo CPU P8400, 2.27GHz in an Hp Pavilion dv4 computer. Figure 3 (a-d) shows the plots of the proportions of transcripts associated with markers for four of the eight traits. There is too much information obtained from the eQTL analysis. Figure S2 (e-h) shows the plots of the remaining four traits. Here, we used grain protein and yield traits as examples to describe the plots. For the grain protein trait, three chromosomes (2, 3, 5) seem to control more genes than other chromosomes. For example, the central region of chromosome 2 contains almost 50% of the 888 transcripts. This region is considered as a hot spot. For the yield trait, chromosome 3 is the only one containing more transcripts. The hot spot is located in the middle of the chromosome and it controls the expression of about 80% of the 457 transcripts.
Other information about this eQTL analysis is provided in Table S4 (Sheet1-Sheet8) and Table S5 (Sheet1-Sheet8). The additional information includes the eQTL effects for each transcript across the genome (Table S5). The posterior probability of each transcript associated with each marker (Table S4). The supplemental tables can serve as a reference database for barley biologists to further study the gene networks for the eight quantitative traits.

4 DISCUSSION

We adopted a new statistical method (SEM) for quantitative trait associated microarray data analysis. We used the method to analyze 22840 microarrayed genes associated with eight quantitative traits in barley. Many genes have been identified to be associated with these traits. The actual functions of these genes in barley are not known prior to this study. These genes are provided in Table S2 (Sheet1-Sheet8) and Table S3. For example, among the 22840 genes, 888 are related to grain protein content. This dataset provides much information for barley biologists to further study these genes. The functions of some genes are known in rice. For example, according to BLASTX results, transcript 22767 (rbah35f01_s_at) is the code for the Cyclopropane-fatty-acyl-phospholipid synthase in rice. This gene is strongly associated to the grain protein in barley, with an F-test statistic of 1436.718 and a p-value of zero.

The same SEM algorithm for phenotype associated microarray data analysis has been applied to eQTL mapping with virtually no modification. The eQTL mapping conducted was still an interval mapping approach where one marker is analyzed at a time. However, all the transcripts were analyzed simultaneously. This is already a significant improvement over the traditional interval mapping for QTL where one transcript was analyzed at a time (Kendziorski, et al., 2006). Results of the eQTL analysis are provided in Table S4 (Sheet1-Sheet8) and Table S5 (Sheet1-Sheet8). This dataset will help barley biologists to infer gene networks for these quantitative traits. Transcripts simultaneously associated with one marker belong to the same network (or pathway) because they are all controlled by the segregation of the same locus. For example, marker ABC152D (83.1 cM) on chromosome 2 controls the expression of about 50% of the transcripts associated with the grain protein. These transcripts are allegedly to be in the same pathway for the development of grain protein.

Fig. 3. Proportions of transcripts associated with markers for the first four traits (Amylase, Diastatic power, Grain protein and Heading date). The chromosomes are separated by the dotted vertical reference lines.
Theoretically, the method can analyze all markers simultaneously using a single model. This is the all-transcript-and-all-marker model. Practically, however, it is difficult to handle the large matrix with a dimensionality repeatedly in the SEM algorithm. The theory is identical to the joint analysis of all the eight traits (a matrix). Further study on this simultaneous analysis is needed for the SEM algorithm. The MCMC implemented Bayesian eQTL mapping (Jia and Xu, 2007) can be adopted here, but it is a sampling based method and it is time consuming in terms of computing time. This study focused on the SEM algorithm for phenotype associated microarray analysis with the eQTL mapping as an example of extension to other problems.

Finally, the data were analyzed using an R program, which can be downloaded from our laboratory website (www.statgen.ucr.edu) under the "Phenotype Associated Microarray" section. A sample dataset (subset of the barley data) is also provided in the website for interested users to test the method. Users may customize the code to analyze their own data using the SEM algorithm.

ACKNOWLEDGEMENTS

Funding: This project was supported by the Agriculture and Food Research Initiative (AFRI) of the USDA National Institute of Food and Agriculture under the Plant Genome, Genetics and Breeding Program 2007-35300-18285 to SX.

REFERENCES


