BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 1 2011, pages 63–69
doi:10.1093/bioinformatics/btq558

Gene expression

Advance Access publication October 29, 2010

A stochastic expectation and maximization algorithm for detecting quantitative trait-associated genes

Haimao Zhan[1], Xin Chen[2] and Shizhong Xu[1,*]

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

Associate Editor: Joaquin Dopazo

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. 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 at *Bioinformatics* online.

Received on June 8, 2010; revised on August 31, 2010; accepted on September 28, 2010

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 ordered phenotype so that a differential expression analysis or an analysis of variances (ANOVA) method can be applied (Cui *et al*., 2005; Kerr *et al*., 2000; Wernisch *et al*., 2003; Wolfinger *et al*., 2001). It is obvious that such discretization is subject to information loss. The current microarray data analysis technique has not been able to efficiently analyze the association of gene expression with a continuous phenotype (Blalock *et al*., 2004; Jia and Xu 2005). Pearson correlation between gene expression and a quantitative trait has been proposed (Blalock *et al*., 2004; Kraft *et al*., 2003; Quackenbush, 2001). Blalock *et al*. (2004) ranked genes according to the correlation coefficients of gene expression with MMSE, a quantitative measurement of the severity of Alzheimer’s disease (AD), and detected many genes that are associated with AD. Kraft *et al*. (2003) used a within family correlation analysis to remove the effect of family stratification. Pearson correlation is intuitive and easy to calculate. However, it may not be optimal because (i) the correlation coefficient may not be the best indicator of the association; (ii) higher order association cannot be detected; (iii) data are analyzed individually with one gene at a time; and (iv) the method cannot be extended to association of gene expression with multiple continuous phenotypes. Potokina *et al*. (2004) investigated the association of gene expression with six malting quality phenotypes (quantitative traits) of 10 barley cultivars. They compared the distance matrix of each gene expression among the 10 cultivars with the distance matrix calculated from the phenotypes of all six traits using the G-test statistic. The G-test statistic was designed to measure the similarity between two matrices. For each gene, there is a distance matrix (based on the expression levels). For the phenotypes of six traits, there is another distance matrix. The two matrices are compared for the similarity. If the similarity is high, the gene is associated with all the six phenotypes. Eventually, the associations of the phenotypes with all the genes are evaluated. The distance matrix comparison approach may have the same flaws as the correlation analysis.

Recently, we proposed to use the regression coefficient of the expression on a continuous phenotype as the indicator of the strength of association (Jia and Xu, 2005). Instead of analyzing one gene at a time, we took a model-based clustering approach to studying all genes simultaneously. Qu and Xu (2006) extended the model-based clustering algorithm to capture genes having higher order association with the phenotype. The model-based clustering analysis (Jia and Xu, 2005; Qu and Xu, 2006) classifies genes into several clusters and all clusters share the same variance–covariance structure. The analysis is implemented via the expectation–maximization (EM) algorithm (Dempster *et al*., 1977). Several 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

*To whom correspondence should be addressed.

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 expression quantitative trait locus mapping (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.pw.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 22 840 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, USA). 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

$$Z_k = 2\frac{X_k - X_{\min}}{X_{\max} - X_{\min}} - 1 \tag{1}$$

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_j be the jth 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 a $n\times p$ 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 (i) Y the microarray data, (ii) Z the phenotypic data and (iii) 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,

$$y_j = X\beta_j + Z\gamma_j + \varepsilon_j \tag{2}$$

where β_j is a $p\times 1$ vector for the effects of cofactors, γ_j is a $q\times 1$ vector for the regression coefficients of gene j on all the q traits. The residual error ε_j is an $n\times 1$ vector with an assumed $N(0, I\sigma^2)$ 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 β_j is

$$\beta_j \sim N(\mu_\beta, \Sigma_\beta) \tag{3}$$

where μ_β is a $p\times 1$ vector of mean and Σ_β is an unknown $p\times p$ positive definite variance matrix. The prior distribution of γ_j is a Gaussian mixture with two components,

$$\gamma_j \sim \pi N(0,\Sigma_1) + (1-\pi)N(0,\Sigma_0) \tag{4}$$

In the above Gaussian mixture, $\Sigma_0 = \omega I_q$ is a known diagonal matrix with a common $\omega = 10^{-8}$ across all the diagonal elements. In other words, Σ_0 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, Σ_1 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 π $(0 < \pi < 1)$ is a prior probability that a gene randomly selected from the pool belongs to cluster 1. A gene classified into cluster one 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

$$\theta = \left\{\mu_\beta, \Sigma_\beta, \Sigma_1, \sigma^2, \pi\right\} \tag{5}$$

We are also interested in ρ_j, the posterior probability of gene j being associated with the traits. The relationship between the posterior ρ_j and the prior π will be presented later.

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

$$\delta_j = \begin{cases} 1 & \text{if } j \text{ belongs to cluster 1} \\ 0 & \text{if } j \text{ belongs to cluster 0} \end{cases} \tag{6}$$

Given δ_j, the genetic effect γ_j has the following distribution,

$$\gamma_j \sim \delta_j N(0,\Sigma_1) + (1-\delta_j)N(0,\Sigma_0) \tag{7}$$

Note that this conditional distribution is not a Gaussian mixture because the membership is already known. The indicator variable δ_j tells which Gaussian component γ_j belongs to. Also note that the parameter vector does not include β_j and γ_j. Under the random model framework, β_j and γ_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

$$p(y_j|\theta,\delta_j) = N\left\{y_j|X\mu_\beta, X\Sigma_\beta X^T + Z\Theta_j Z^T + I\sigma^2\right\} \tag{8}$$

where

$$\Theta_j = \delta_j\Sigma_1 + (1-\delta_j)\Sigma_0 \tag{9}$$

and the notation $N\{y|a,b\}$ stands for the normal density of variable y with mean a and variance b. Given the cluster membership, $\delta = \left\{\delta_j\right\}$, the log likelihood function for the entire dataset is

$$L(\theta|\delta)=\sum_{j=1}^{m}\ln\left[p(y_j|\theta,\delta_j)\right] \tag{10}$$

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,

$$p_1(y_j|\theta)=p(y_j|\theta,\delta_j=1)=N(y_j|X\mu_\beta,X\Sigma_\beta X^T+Z\Sigma_1 Z^T+I\sigma^2) \tag{11}$$

and

$$p_0(y_j|\theta)=p(y_j|\theta,\delta_j=0)=N(y_j|X\mu_\beta,X\Sigma_\beta X^T+Z\Sigma_0 Z^T+I\sigma^2) \tag{12}$$

The posterior probability that $\delta_j=1$ is

$$\rho_j=\mathrm{E}(\delta_j|\theta,y_j)=\frac{\pi p_1(y_j|\theta)}{\pi p_1(y_j|\theta)+(1-\pi)p_0(y_j|\theta)} \tag{13}$$

Because δ_j is a Bernoulli variable, it is sampled from

$$p(\delta_j)=\mathrm{Bernoulli}(\delta_j|\rho_j) \tag{14}$$

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 γ_j conditional on δ_j by

$$\Theta_j=var(\gamma_j|\delta_j)=\delta_j\Sigma_1+(1-\delta_j)\Sigma_0 \tag{15}$$

Let us define

$$V_j=X\Sigma_\beta X^T+Z\Theta_j Z^T+I\sigma^2 \tag{16}$$

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

$$\pi=\frac{1}{m}\sum_{j=1}^{m}\delta_j \tag{17}$$

The population mean μ_β is updated using

$$\mu_\beta=\left[\sum_{j=1}^{m}X^T V_j^{-1}X\right]^{-1}\left[\sum_{j=1}^{m}X^T V_j^{-1}y_j\right] \tag{18}$$

The variance-covariance matrix of β_j is denoted by Σ_β and updated using

$$\Sigma_\beta=\frac{1}{m}\sum_{j=1}^{m}E(\beta_j\beta_j^T)=\frac{1}{m}\sum_{j=1}^{m}\left[E(\beta_j)E(\beta_j^T)+var(\beta_j)\right] \tag{19}$$

where

$$\mathrm{E}(\beta_j)=\Sigma_\beta X^T V_j^{-1}(y_j-X\mu_\beta) \tag{20}$$

and

$$var(\beta_j)=\Sigma_\beta-\Sigma_\beta X^T V_j^{-1}X\Sigma_\beta \tag{21}$$

Given δ_j, the unknown variance–covariance matrix of γ_j is $\Theta_j=\delta_j\Sigma_1+(1-\delta_j)\Sigma_0$. However, the corresponding matrix Σ_0 is a constant. Therefore, we only need to update Σ_1 using all γ_j that come from cluster one. The updated equation for Σ_1 is

$$\Sigma_1=\frac{1}{\pi m}\sum_{j=1}^{m}\delta_j E(\gamma_j\gamma_j^T)=\frac{1}{\pi m}\sum_{j=1}^{m}\delta_j\left[E(\gamma_j)E(\gamma_j^T)+var(\gamma_j)\right] \tag{22}$$

where

$$\mathrm{E}(\gamma_j)=\Sigma_1 Z^T V_j^{-1}(y_j-X\mu_\beta) \tag{23}$$

and

$$var(\gamma_j)=\Sigma_1-\Sigma_1 Z^T V_j^{-1}Z\Sigma_1 \tag{24}$$

The residual error variance is updated using

$$\sigma^2=\frac{1}{mn}\sum_{j=1}^{m}\left(y_j-X\mu_\beta\right)^T\left(y_j-X\mu_\beta-XE(\beta_j)-\delta_j ZE(\gamma_j)\right) \tag{25}$$

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 $\theta=\{\mu_\beta,\Sigma_\beta,\Sigma_1,\sigma^2,\pi\}$. So far all parameter have been updated. We can now combine the stochastic steps with the EM steps to compete 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 $\delta=\{\delta_j\}$. We can monitor the converging process for each parameter. Once all parameters have converged, we start to collect the posterior sample for θ. 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 $\theta^{(t)}=\left\{\mu_\beta^{(t)},\Sigma_\beta^{(t)},\Sigma_1^{(t)},\sigma^{2(t)},\pi^{(t)}\right\}$ be the t-th observation in the posterior sample (after convergence), the estimate parameter vector of the SEM algorithm is

$$\hat{\theta}=\frac{1}{T}\sum_{t=1}^{T}\theta^{(t)} \tag{26}$$

The most important quantity of the SEM analysis is not the entire vector of θ; rather, it is

$$\hat{\rho}_j=\frac{1}{T}\sum_{t=1}^{T}\rho_j^{(t)}\approx\frac{1}{T}\sum_{t=1}^{T}\delta_j^{(t)} \tag{27}$$

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 $\rho_j\geq 0.9$ to declare significant association.

2.6 Expression quantitative trait locus 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

$$y_j = X\beta_j + Z_k\gamma_{jk} + \varepsilon_j \tag{28}$$

where γ_{jk} is the QTL effect for transcript j at marker k. The Z_k variable is defined as

$$Z_k = \begin{cases} +1 & \text{for } A_1A_1 \\ -1 & \text{for } A_2A_2 \end{cases} \tag{29}$$

where A_1A_1 is first genotype and A_1A_1 is the second genotype of marker k. The barley population under study is a doubled haploid population and thus only two genotypes exist for each locus.

We now have to analyze the data Q times, one for each marker. Previously, we have a single π for the proportion of genes associated with the phenotype. We now have Q suchπ's to indicate the proportions of transcripts associated with all markers, denoted by a vector

$$\pi = [\,\pi_1 \ \cdots \ \pi_Q\,] \tag{30}$$

In addition, the posterior probability of gene j associated with marker k is denoted by ρ_{jk}. These parameters $\{\pi_k, \rho_{jk}\}$ 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 $m = 22\,840$. Since no other cofactor existed except the intercept, β_j is a scalar with dimension $p = 1$. For the single trait analysis, $q = 1$ for each trait analysis. Each of the five parameters $\theta = \{\,\mu_\beta,\ \Sigma_\beta,\ \Sigma_1,\ \sigma^2,\ \pi\,\}$ 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 from the separate SEM analyses are listed in Supplementary Table S1. Lists of all the detected genes associated with the traits and gene annotations are given in Supplementary Table S2 (Sheet1–Sheet8) for interested readers.

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

Trait	SEM algorithm		EM algorithm	
	Number	Proportion (%)	Number	Proportion (%)
Alpha-amylase	644	2.82	233	1.02
Diastatic power	467	2.04	195	0.85
Grain protein	888	3.89	257	1.13
Grain yield	457	2.00	139	0.61
Heading date	401	1.76	166	0.73
Height	784	3.43	264	1.16
Lodging	650	2.85	173	0.76
Malt extract	526	2.30	216	0.95

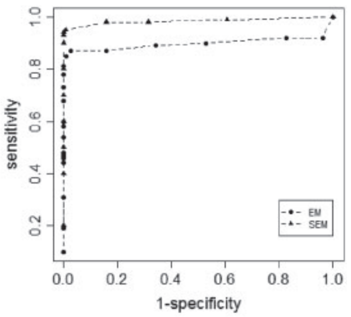

Fig. 1. ROC curve comparing the SEM and EM algorithms in the simulation study.

In order to compare the performance of the SEM model and the EM model, we carried out a simulation experiment based on the SEM model by using the estimated parameters obtained from the barley data (grain yield) analysis. The ROC curve (Fig. 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 EM 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

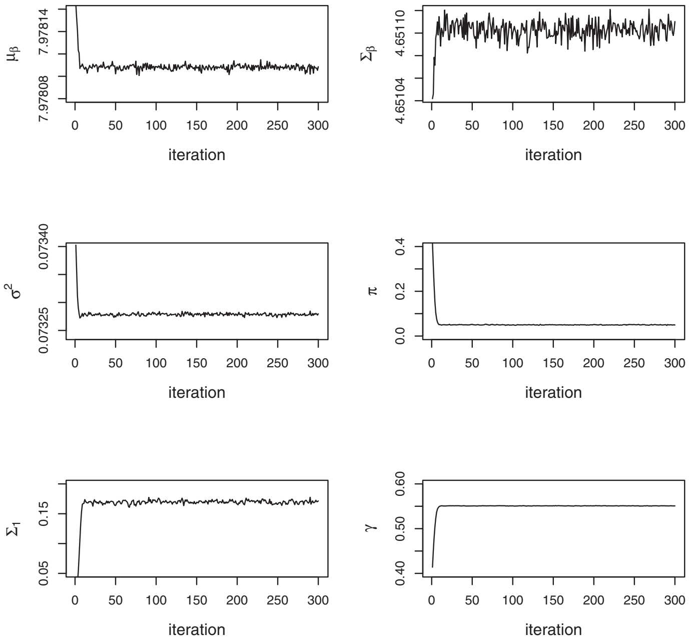

Fig. 2. The SEM convergence processes of five parameters for the grain yield trait and the regression coefficient (γ) of gene AF250937_s_at (the gene having strong association with grain yield).

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 (Supplementary Figure S1). Some genes have positive associations with the traits, e.g. AF250937_s_at and Contig6445_at, and some have negative associations with the traits, e.g. Contig23592_at and Contig3295_at.

3.2 Multiple traits association study

For the joint association study of eight traits, the dimensionality of the parameters increased to $p=8$ and $q=8$. Therefore, μ_β is an 8×1 vector, Σ_β is a 8×8 matrix and Σ_1 is an 8×8 matrix. The rest of the parameters, π and σ^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 $\rho_j \ge 0.9$ 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 22 840 genes included in the analysis. The list of associated genes is given in the supplemental data (Supplementary Table S3) for interested readers. The estimated parameters are $\hat{\pi}=0.089$ and $\hat{\sigma}^2=0.067$ for the proportion and residual error variance, respectively. The estimated μ_β and Σ_β are $\hat{\mu}_\beta=7.978$ and $\hat{\Sigma}_\beta=4.652$, respectively. The estimated variance matrix of the differentially expressed cluster is

$$\hat{\Sigma}_1=$$

$$\begin{bmatrix} 0.107 & 0.002 & -0.027 & 0.023 & 0.027 & -0.074 & 0.033 & 0.023 \\ 0.002 & 0.052 & -0.026 & 0.009 & 0.007 & 0.002 & -0.013 & 0.005 \\ -0.027 & -0.026 & 0.204 & -0.095 & 0.0003 & 0.017 & -0.026 & -0.012 \\ 0.023 & 0.009 & -0.095 & 0.084 & 0.015 & -0.009 & 0.026 & -0.008 \\ 0.027 & 0.007 & 0.0003 & 0.015 & 0.096 & 0.039 & -0.025 & -0.024 \\ -0.074 & 0.002 & 0.017 & -0.009 & 0.039 & 0.198 & -0.097 & -0.048 \\ 0.033 & -0.013 & -0.026 & 0.026 & -0.025 & -0.097 & 0.193 & -0.054 \\ 0.023 & 0.005 & -0.012 & -0.008 & -0.024 & -0.048 & -0.054 & 0.151 \end{bmatrix}$$

Note that the proportion of genes detected (7.2%) in the experiment is not the same as $\hat{\pi}=8.9\%$ because the former depends on the cut-off point ($\rho_j \ge 0.9$) 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 22 840 genes, 521 genes still had significant effects in the permuted data. The false positive is 521/22 840 = 0.022 811, 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

$$\hat{F}_j=\frac{1}{q}\hat{\gamma}_j^T\left(\hat{\Sigma}_1-\hat{\Sigma}_1 Z^T\hat{V}_j^{-1}Z\hat{\Sigma}_1\right)^{-1}\hat{\gamma}_j \tag{31}$$

where,

$$\hat{\gamma}_j=\hat{\Sigma}_1 Z^T\hat{V}_j^{-1}(y_j-X\hat{\mu}_\beta) \tag{32}$$

and

$$\hat{V}_j=X\hat{\Sigma}_\beta X^T+Z\hat{\Sigma}_1 Z^T+I\hat{\sigma}^2 \tag{33}$$

The P-value was calculated using

$$P\text{-value}=1-F_{8,\infty}(\hat{F}_j) \tag{34}$$

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

3.3 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 22 840 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

Table 2. The estimated regression coefficients for the top 10 genes jointly associated with all eight traits

Rank	Probe set ID	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7	γ_8	F-test	P-value
1	Contig1132_s_at	−2.99	0.99	3.42	−1.99	−0.25	0.75	−0.40	−0.05	945.744	<0.0001
2	Contig124_at	2.32	−0.68	−2.93	1.76	0.38	−0.55	0.41	0.12	581.689	<0.0001
3	Contig2163_at	−0.12	−0.77	−1.88	0.62	−0.89	0.45	2.15	1.62	543.998	<0.0001
4	Contig11524_at	0.71	−0.83	1.18	−0.10	2.30	0.61	1.62	−1.63	492.936	<0.0001
5	Contig2279_at	0.10	0.35	−0.72	0.80	−0.82	−2.05	−1.51	4.94	481.062	<0.0001
6	Contig4769_at	2.67	−0.18	−2.32	1.23	0.55	−0.82	0.28	0.21	435.547	<0.0001
7	Contig23772_s_at	1.81	−0.67	−2.53	1.50	0.13	−0.55	0.59	0.003	421.314	<0.0001
8	HVSMEf0020F06r2_at	0.55	−0.74	0.92	−0.16	2.16	0.75	1.33	−1.45	396.722	<0.0001
9	Contig11126_at	−0.44	−0.21	0.62	−0.86	0.88	1.91	1.09	−4.10	348.061	<0.0001
10	Contig4526_at	−0.45	−1.17	1.11	0.88	1.27	1.38	0.58	−1.35	347.999	<0.0001

The partial regression coefficients of gene on the eight traits are denoted by $\gamma_1,...,\gamma_8$.

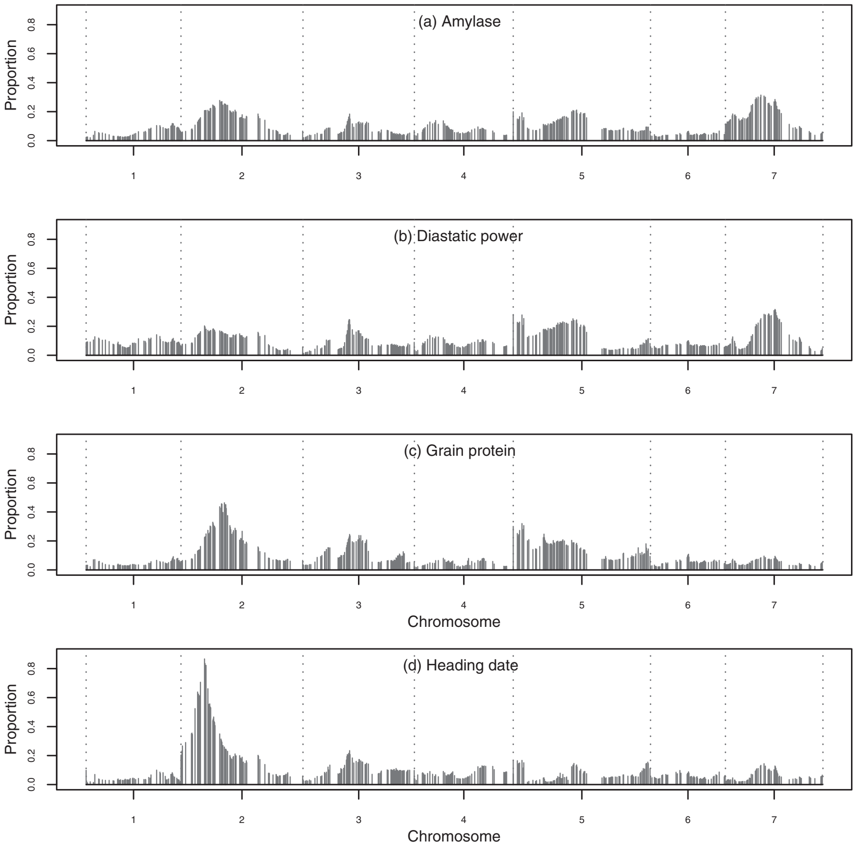

Fig. 3. Proportions of transcripts associated with markers for the first four traits (Amylase, Diastatic power, Grain protein and Heading date). The chromosomes are separated by the dotted vertical reference lines.

average marker interval <2 centiMorgan, covering the entire barley genome.

The entire eQTL analysis took about nine hours of Intel Core Duo CPU P8400, 2.27 GHz in an Hp Pavilion dv4 computer. Figure 3a–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. Supplementary Figure S2e–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 Supplementary Table S4 (Sheet1–Sheet8) and Supplementary

Table S5 (Sheet1–Sheet8). The additional information includes the eQTL effects for each transcript across the genome (Supplementary Table S5). The posterior probability of each transcript associated with each marker (Supplementary Table S4). The supplementary 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 22 840 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 Supplementary Table S2 (Sheet1–Sheet8) and Supplementary Table S3. For example, among the 22 840 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 22 767 (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 Supplementary Table S4 (Sheet1–Sheet8) and Supplementary 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.

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 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.

Funding: 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 S.X.).

Conflict of Interest: none declared.

REFERENCES

Blalock,E.M. *et al*. (2004) Incipient Alzheimer’s disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proc. Natl Acad. Sci. USA*, **101**, 2173–2178.

Celeux,G. and Diebolt,J. (1986) The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Comput. Statist. Quart*., **2**, 73–82.

Cookson,W. *et al*. (2009) Mapping complex disease traits with global gene expression. *Nat. Rev. Genet*., **10**, 184–194.

Cui,X. *et al*. (2005) Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, **6**, 59–75.

Dempster,A.P. *et al*. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc*., **39**, 1–38.

Hayes,P.M. *et al*. (1993) Quantitative trait locus effects and environmental interaction in a sample of North American barley germ plasm. *Theor. Appl. Genet*., **87**, 392–401.

Irizarry,R.A. *et al*. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Jia,Z. and Xu,S. (2005) Clustering expressed genes on the basis of their association with a quantitative phenotype. *Genet Res*., **86**, 193–207.

Jia,Z. and Xu,S. (2007) Mapping quantitative trait loci for expression abundance. *Genetics*, **176**, 611–623.

Kendziorski,C.M. *et al*. (2006) Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics*, **62**, 19–27.

Kerr,M.K. *et al*. (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol*., **7**, 819–837.

Kraft,P. *et al*. (2003) A family-based test for correlation between gene expression and trait values. *Am. J. Hum. Genet*., **72**, 1323–1330.

Lander,E.S. and Botstein,D. (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **121**, 185–199.

Luo,Z.W. *et al*. (2007) SFP genotyping from affymetrix arrays is robust but largely detects cis-acting expression regulators. *Genetics*, **176**, 789–800.

Morley,M. *et al*. (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430**, 743–747.

Potokina,E. *et al*. (2004) Functional association between malting quality trait components and cDNA array based expression patterns in barley (Hordeum vulgare L.). *Mol. Breeding*, **14**, 153–170.

Qu,Y. and Xu,S. (2006) Quantitative trait associated microarray gene expression data analysis. *Mol. Biol. Evol*., **23**, 1558–1573.

Quackenbush, J. (2001) Computational analysis of microarray data. *Nat. Rev. Genet*., **2**, 418–427.

Wernisch,L. *et al*. (2003) Analysis of whole-genome microarray replicates using mixed models. *Bioinformatics*, **19**, 53–61.

Wolfinger,R.D. *et al*. (2001) Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol*., **8**, 625–637.