www.nature.com/hdy

Bayesian mapping of genotype × expression interactions in quantitative and qualitative traits

F Hoti and MJ Sillanpää

Department of Mathematics and Statistics, Rolf Nevanlinna Institute, University of Helsinki, FIN-00014 Helsinki, Finland

A novel Bayesian gene mapping method, which can simultaneously utilize both molecular marker and gene expression data, is introduced. The approach enables a quantitative or qualitative phenotype to be expressed as a linear combination of the marker genotypes, gene expression levels, and possible genotype × gene expression interactions. The interaction data, given as marker–gene pairs, contains possible *in cis* and *in trans* effects obtained from earlier allelic expression studies, genetical genomics studies, biological hypotheses, or known pathways. The method is presented for an inbred line cross design and can be easily generalized to handle other types of populations and designs. The model selection is based on the use of effect-specific variance components combined with Jeffreys' non-informative prior – the method operates by adaptively shrinking marker, expression, and interaction effects toward zero so that non-negligible effects are expected to occur only at very few positions. The estimation of the model parameters and the handling of missing genotype or expression data is performed via Markov chain Monte Carlo sampling. The potential of the method including heritability estimation is presented using simulated examples and novel summary statistics. The method is also applied to a real yeast data set with known pathways.

Heredity (2006) **97,** 4–18. doi:10.1038/sj.hdy.6800817; published online 3 May 2006

Keywords: gene expression; QTL; molecular markers; sparse method; Bayes

Introduction

Population-based association analysis, enhanced with projects like HAPMAP (The International HapMap Consortium, 2003, 2005), is one of the most promising gene mapping techniques (Risch and Merikangas, 1996; Lohmueller *et al*, 2003). The selection of a trait-associated subset of markers among a large number of candidates is a challenging model selection problem on its own right (Broman and Speed, 2002; Sillanpää and Corander, 2002; Kilpikari and Sillanpää, 2003). In addition to measuring the genotype at selected marker points along the chromosome, it is currently possible to measure the gene expression, mRNA abundance, levels for a large number of genomic locations simultaneously. The availability of such data has made it possible to base candidate selection on the associations between a phenotype and gene expression levels (Quackenbush, 2001; Wayne and McIntyre, 2002; Kraft *et al*, 2003; Goeman *et al*, 2004; Lu *et al*, 2004). Further, if both phenotype–marker and phenotype–expression associations are analyzed, it is possible to study the overlap between the genomic locations of the resulting significant markers and genes (Aune *et al*, 2004).

Gene expression and marker data are combined in genetical genomics – to map gene regulators or modifier genes with respect to a marker map. This is carried out by treating the expression levels as phenotypes in a similar manner as quantitative traits in standard quantitative trait locus (QTL) mapping (for alternative procedures, see Jansen and Nap, 2001; Jansen, 2003; Darvasi, 2003; Heighway *et al*, 2005). Such practice has led to accumulating evidence that genetic variants can control or introduce quantitative variation in gene expression levels, and that this variation follows the Mendelian inheritance similar to the genetic variants themselves (Brem *et al*, 2002; Yan *et al*, 2002; Lo *et al*, 2003; Schadt *et al*, 2003; Jansen and Nap, 2004; Knight, 2004; Morley *et al*, 2004; Auger *et al*, 2005; Bystrykh *et al*, 2005; Chesler *et al*, 2005; Hubner *et al*, 2005).

The treatment of gene expression levels as genetic markers, expression level polymorphisms, in QTL studies has been suggested by Doerge (2002) and Kell (2002). The utilization of gene expression profiling to define genetically homogeneous groups or to improve the definition of a phenotype for gene mapping purposes was proposed by Watts *et al* (2002), Schadt *et al* (2003), and Kraft and Horvath (2003). Instead of studying each expression phenotype individually in genetical genomics, Perez-Enciso *et al* (2003) addressed the problem of combining gene expression phenotypes and QTL mapping by first predicting the value of an underlying liability variable using partial least squares. In their approach, the hypothetical liability (underlying the observed phenotype) is defined as a linear combination of individual gene expression levels. This predicted liability is then used instead of the individual expression phenotypes in QTL mapping. The use of principal component analysis and hierarchical clustering for a similar purpose has been proposed by Lan *et al* (2004).

Correspondence: MJ Sillanpää, Department of Mathematics and Statistics, Rolf Nevanlinna Institute, University of Helsinki, PO Box 68, FIN-00014 Helsinki, Finland. E-mail: mjs@rolf.helsinki.fi

Received 9 September 2005; accepted 28 February 2006; published online 3 May 2006

The large size of modern gene expression and molecular marker data sets combined with the goal of finding a small subset of trait-associated candidate genes underlines the need for computationally efficient methods especially designed to detect sparse candidates. Recently, sophisticated shrinkage and sparse methods have been proposed to study phenotype–expression association (Shevade and Keerthi, 2003; West, 2003; Lopes and West, 2004) and phenotype–marker association (Meuwissen *et al*, 2001; Devlin *et al*, 2003; Kopp *et al*, 2003; Xu, 2003; Sillanpää and Bhattacharjee, 2005; Zhang and Xu, 2005; Zhang *et al*, 2005).

In this paper, we present a novel method where the phenotype is modeled as a linear combination of marker genotypes, gene expression levels, and genotype × gene expression interaction terms. The method is implemented for inbred line cross data by using a Bayesian shrinkage approach in a similar manner as in Meuwissen *et al* (2001) and Xu (2003). Our method is computationally efficient and easy to use, since no tuning parameters are required. Also similar to Xu (2003), our method provides good heritability estimates. We emphasize the use of standardized regression coefficients in interpreting the results and introduce summary statistics that enable us to effectively identify complementary genetic determinants (submodels).

Model

Genetic model

Here, we consider a population with only two segregating genotypes resulting from an inbred line cross experiment, for example a backcross or a double haploid population. The generalization to situations with multiple genotypes, that is, general population samples of outbred human populations, is also discussed. We assume that the sample consists of both marker data and gene expression measurements (mRNA abundance levels), as well as a quantitative or a qualitative trait phenotype, that have been collected from all study individuals. In addition, we assume that a proportion of the marker measurements are *a priori* associated with some of the gene expression measurements, allowing the identification of genotype-specific gene expression effects, that is, genotype × expression interactions with respect to the phenotype. We allow multiple markers to be associated with a single gene and *vica versa*. In summary, our genetic data consist of three data subtypes of the following forms:

1. marker data (N_M markers),
2. gene expression data (N_E genes),
3. link data: data allowing the identification of marker–expression pairs whose interaction term is to be considered in our model (N_{ME} marker–gene pairs).

The gene expression data are assumed to have gone through suitable transformation and normalization steps (Quackenbush, 2001; Butte, 2002) so that the sample distribution of the majority of the genes is approximately standard normal. The link data, which can originate from previous genetical genomics studies (*cis*- and *trans*-acting variation) or are based on known pathways, enable incorporating cross terms into the model. If no prior external knowledge or hypothesis is available, the link data can be constructed solely based on the genetic distances between the markers and the genes. Thus, one assumes *in cis* effects between the marker and all genes within a given genetic distance of the marker. Also, oligonucleotide arrays can provide simultaneous genotype and gene expression measurements directly (Ronald *et al*, 2005).

Given genetic data as described above, we propose modeling a quantitative phenotype y_i of individual i with the following linear model:

$$\begin{aligned} y_i = \mu &+ \sum_{j=1}^{N_M}\sum_{k=1}^{2}\alpha_{j,k}z_{i,j,k} + \sum_{j=1}^{N_E}\beta_j x_{i,j} \\ &+ \sum_{j=1}^{N_{ME}}\sum_{k=1}^{2}\gamma_{j,k}\tilde{z}_{i,j,k}\tilde{x}_{i,j} + \varepsilon_i \end{aligned} \qquad (1)$$

where μ is the population mean and $\varepsilon_i \sim N(0, \sigma_0^2)$ is a normally distributed residual term with mean zero and variance σ_0^2. For a binary trait, we use model (1) to model an underlying continuous liability, which then gives rise to the binary observation according to the Bayesian probit model (see Appendix A1 for details). The first summation runs over all markers and is designed to capture the genotype-specific main effects (cf Xu, 2003). For individual i at marker j, the indicator for genotype k is denoted by $z_{i,j,k}$, and $\alpha_{j,k}$ is the coefficient of the corresponding genotype-specific main effect. The second summation runs over the N_E genes and $x_{i,j}$ denotes the gene expression measurement of gene j for individual i, and β_j is the coefficient of the corresponding linear gene expression effect. In the last summation, genotype and gene expression data of the N_{ME} marker–gene pairs are gathered into pairs $\{\tilde{x}_{i,j}, \tilde{z}_{i,j,k}\}$, $j = 1, \ldots, N_{ME}$, where $\tilde{x}_{i,j} = x_{i,g_j}$ and $\tilde{z}_{i,j,k} = z_{i,s_j,k}$ for some pair (g_j, s_j) given by the link data. Thus, for individual i, $\tilde{x}_{i,j}$ is the gene expression measurement of gene g_j and $\tilde{z}_{i,j,k}$ is the indicator of genotype k for the corresponding marker s_j, and $\gamma_{j,k}$ is the coefficient of the corresponding genotype × expression interaction effect. The extension of model (1) to include also genotype × genotype or expression × expression interaction terms is considered in the Discussion section.

In order to ensure that the model parameters are identifiable, we introduce the constrains

$$\alpha_{j,1} = 0 \quad \text{for } j = 1, \ldots, N_M \qquad (2)$$

$$\gamma_{j,1} = 0 \quad \text{for } j = 1, \ldots, N_{ME} \qquad (3)$$

Thus, the first genotype, at each marker, is identified as a baseline, and their effects are included into the terms μ and $\sum_j \beta_j x_{i,j}$. The genotype-specific contrasts (differences) are then modeled using the two remaining terms in model (1), which, by taking into account the above constrains, are $\sum_j \alpha_{j,2}z_{i,j,2}$ and $\sum_j \gamma_{j,2}\tilde{z}_{i,j,2}\tilde{x}_{i,j}$.

Next, for individual i, the genetic data are gathered into vector

$$\begin{aligned} X_i = &(z_{i,1,2}, \ldots, z_{i,N_M,2}, x_{i,1}, \ldots, x_{i,N_E}, \tilde{z}_{i,1,2}\tilde{x}_{i,1}, \\ &\ldots, \tilde{z}_{i,N_{ME},2}\tilde{x}_{i,N_{ME}}) \end{aligned}$$

and the vector containing the $N = N_M + N_E + N_{ME}$ unknown effects (coefficients) is denoted by

$$\theta = (\alpha_{1,2}, \ldots, \alpha_{N_M,2}, \beta_1, \ldots, \beta_{N_E}, \gamma_{1,2}, \ldots, \gamma_{N_{ME},2})$$

Now, by taking into account constrains (2) and (3), the linear model (1) can be rewritten as

$$y_i = \mu + \sum_{j=1}^{N} \theta_j X_{i,j} + \varepsilon_i \tag{4}$$

Hierarchical model

Prior distributions: Bayesian approaches require the specification of prior distributions for the unknown parameters. We follow the work of Xu (2003) and adopt the following prior densities, where each effect is assigned its own variance term. For $j=1, \ldots, N$, the effect prior $p(\theta_j \mid \sigma_j^2)$ is the density function of normal distribution with mean zero and effect-specific variance σ_j^2, and $p(\sigma_j^2) \propto 1/\sigma_j^2$ is the (Jeffreys' scale invariant) prior density function of the effect-specific hyperparameter σ_j^2. The prior density function of the mean μ is $p(\mu) \propto 1$, and the prior density function of the variance $\sigma_0^2 = \mathrm{var}(\varepsilon_i)$, for $i=1, \ldots, n$, is $p(\sigma_0^2) \propto 1/\sigma_0^2$. Now, by the use of appropriate (conditional) independence assumptions, the joint prior density function of the model parameters θ, μ, and σ^2, where $\sigma^2 = (\sigma_0^2, \ldots, \sigma_N^2)$, is $p(\theta,\mu,\sigma^2) = p(\theta \mid \sigma^2) p(\mu) p(\sigma^2)$, where $p(\theta|\sigma^2) = \prod_{j=1}^{N} p(\theta_j|\sigma_j^2)$, and $p(\sigma^2) = \prod_{j=0}^{N} p(\sigma_j^2)$. It has been demonstrated (see Figueiredo, 2003; Xu, 2003) that the above use of effect-specific variance parameters induces sparseness. Thus, our prior information states that most terms in the sums of model (1) are expected to be zero or almost zero and the degree of sparseness adaptively depends on the data at hand.

Model for missing values: In Bayesian inference, missing values are handled in a similar manner as any other unknown parameter (random variable). Thus, prior distributions are assigned to all missing values. The prior density function $p(x_{i,j})$ of a missing gene expression measurement $x_{i,j}$ is chosen to be that of a standard normal distribution. Recall that the gene expression level measurements are assumed to be approximately normally distributed.

Next we define the prior distribution for the marker data, in a backcross or a double haploid situation, by taking into account the probability of a recombination, which again is defined by the genetic distances between the markers. Following Sillanpää and Arjas (1998), the joint probability of the marker data for individual i is given by

$$P(m_{i,1}, \ldots, m_{i,N_M}) \propto P(m_{i,1}) \prod_{j=2}^{N_M} P(m_{i,j}|m_{i,j-1}) \tag{5}$$

where $m_{i,j}$ is the genotype of individual i at marker j, $P(m_{i,1}) = \frac{1}{2}$ is the probability of genotype $m_{i,1}$ at marker 1, and $P(m_{i,j} \mid m_{i,j-1})$ is the probability of genotype $m_{i,j}$ at marker j conditional on genotype $m_{i,j-1}$ at marker $j-1$. The conditional probability $P(m_{i,j} \mid m_{i,j-1})$ is $1-r_j$ if genotypes $m_{i,j}$ and $m_{i,j-1}$ are the same (no recombination) and is r_j otherwise, where r_j is derived from the genetic distance d_j (in Morgans) between markers j and $j-1$, by the Haldane map function $r_j = \frac{1}{2}(1 - \exp(-2|d_j|))$. A simpler way to proceed, which works also in more general setups, is to assume independence between markers and take the prior probability of each genotype to be equal. However, in many cases, it is possible to derive more informative prior distributions similar to equation (5); for various crosses from two inbred lines, see Jiang and Zeng (1997).

Posterior distributions: Next we derive the posterior distribution of the model parameters θ, μ, and σ^2, where $\sigma^2 = (\sigma_0^2, \ldots, \sigma_N^2)$. Denote by $D = \{m,x\}$ the complete genetic data, that is, the combined marker and gene expression data, with no missing values. Further, let $D^- = \{m^-, x^-\}$ denote the observed genetic data with possibly some entries missing. By the use of the Bayes formula, the density function of the joint posterior distribution (see Figure 1) of the model parameters and the genetic data is given by

$$\begin{aligned} p(\theta, \mu, \sigma^2, D|y, D^-) \propto{} & p(\theta, \mu, \sigma^2) \\ & \times p(D) p(D^-|D) p(y|\theta, \mu, \sigma,^2 D) \end{aligned} \tag{6}$$

where $p(\theta, \mu, \sigma^2)$ is the density function of the joint prior distribution of the parameters (θ, μ, σ^2), $p(D)$ is the prior density function of the complete genetic data D, $p(D^- \mid D)$ is the mass probability function of the observed genetic data D^- conditional on the complete genetic data D, and $p(y \mid \theta, \mu, \sigma^2, D)$ is the likelihood of the phenotype data y. Note that $p(D^- \mid D)$ is an indicator function and takes value 1 only when D^- is consistent with D and 0 otherwise. The prior density function of the genetic data D is proportional to $\prod_{i=1}^{n} (p(m_{i,1}, \ldots, m_{i,N_M}) \prod_{j=1}^{N_E} p(x_{i,j}))$ and the likelihood function can be factorized into $\prod_{i=1}^{n} p(y_i|\theta, \mu, \sigma^2, D)$, where

$$p(y_i|\theta, \mu, \sigma^2, D) \propto \frac{1}{\sqrt{\sigma_0^2}} \exp\left(-\frac{1}{2\sigma_0^2} \left(y_i - \mu - \sum_{j=1}^{N_M} \alpha_{j,2} z_{i,j,2} - \sum_{j=1}^{N_E} \beta_j x_{i,j} - \sum_{j=1}^{N_{ME}} \gamma_{j,2} \tilde{z}_{i,j,2} \tilde{x}_{i,j} \right)^2 \right)$$

Markov chain Monte Carlo estimation: We apply Gibbs sampling (Geman and Geman, 1984) and Metropolis–Hastings (Hastings, 1970) algorithms to draw dependent samples from the joint posterior distribution of the unknowns (equation (6)). The specific choices of the prior distributions (conjugate distributions) allow us to generate samples directly from the fully conditional marginal posterior distributions of θ, μ, and σ^2. A detailed description of the adopted algorithm is given in Appendix A1. Possible point estimates for the unknown distributions include the maximum a posteriori (MAP), the median, and the expected value of the marginal distributions. We assume that the number of markers and genes in the data set is such that it is computationally reasonable to attempt to estimate the complete posterior distribution, for example, their number is reduced by some preliminary feature selection algorithm or the features are chosen based on known pathways (see Thomas, 2005). Zhang and Xu (2005) were able to handle a model where the number of effects was 15 times larger than the sample size. In our opinion, it is preferable to reduce this ratio (upper limit) even lower, say down to 10, by gathering more samples or by reducing the number of considered effects in the model. If the number of markers or genes is very large (several thousands), even though the data contain enough information to estimate the effects, the

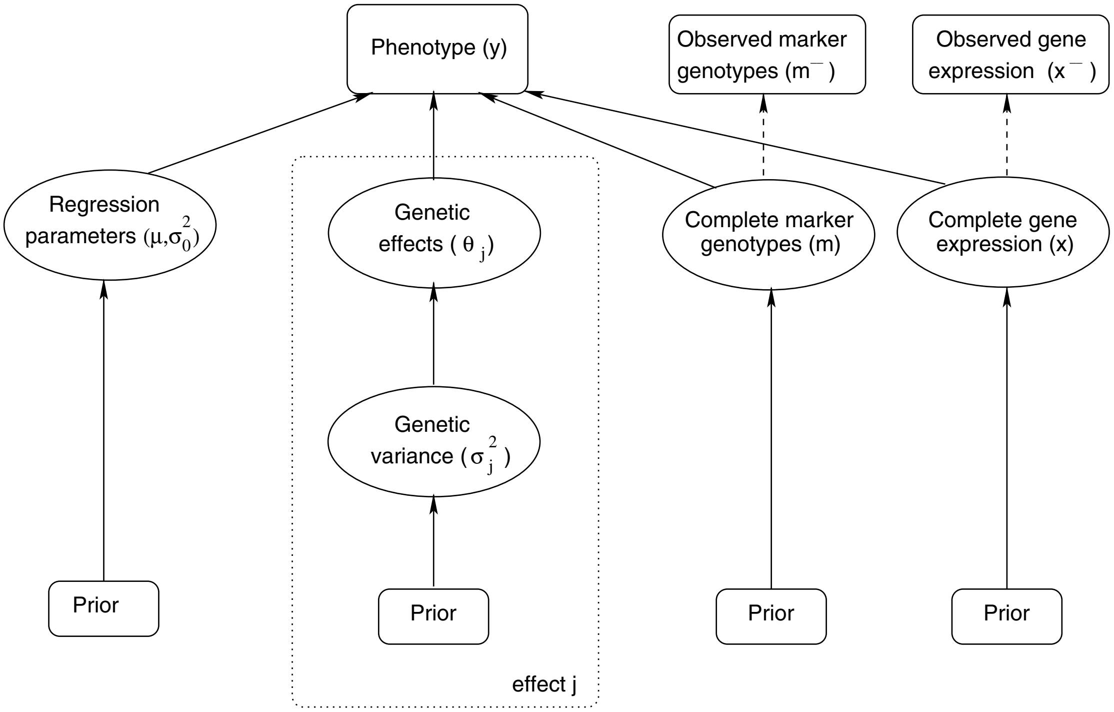
Phenotype (y)
Observed marker genotypes (m^-)
Observed gene expression (x^-)
Regression parameters (μ,σ_0^2)
Genetic effects (θ_j)
Complete marker genotypes (m)
Complete gene expression (x)
Genetic variance (σ_j^2)
Prior
Prior
Prior
Prior
effect j

Figure 1 Graphical representation of the hierarchical structure of the model. Boxes refer to prespecified values or observed data and ellipses to unknown random variables. This graph (directed acyclic graph, DAG) gives a graphical summary of the conditional independence assumptions and directions of hierarchical (solid arrows) and deterministic (dotted arrows) dependencies.

time needed to perform the calculations can be overwhelming. In such case, one can postpone the estimation of the whole distribution and concentrate on directly estimating some summary statistic of the distribution (eg MAP via an EM algorithm in Figueiredo (2003), and MAP via a penalized ML method in Zhang and Xu (2005)).

Simulations

We simulated backcross data consisting of molecular markers, gene expression level measurements, and both a binary and a continuous phenotype. First, linked marker data were simulated, then gene expression data were generated conditionally on the marker data. Next, the phenotype data were generated conditionally on both the marker and gene expression data. Finally, missing values were introduced by randomly removing a given proportion of the marker and gene expression measurements. Note that our simulation strategy differs from some others, which use real gene expression data as a starting point (Perez-Enciso *et al*, 2003; Perez-Enciso, 2004). By conditioning on the expression measurement, Perez-Enciso *et al* (2003) simulated case–control data (QTLs and binary phenotypes) using partial least squares. Subsequent linked marker data were then generated around each QTL using the exponential decay model for linkage disequilibrium. In Perez-Enciso (2004), the data, a set of linked marker loci and phenotypes, were simulated from an outbred population based on coalescent techniques and gene dropping. Our main reason for not adopting any existing simulation method was the need to have realistic linked marker data for offspring resulting from a backcross design of two inbred lines. Also, our approach (see below), although including simplifying assumptions, allows us to fully validate the performance of the proposed estimation method.

Genetic data

Linked marker data for a population of 200 backcross individuals was simulated using the QTL Cartographer software (Basten *et al*, 1994, 2003). Altogether 100 markers were simulated, 50 markers equally spaced on two different chromosomes. The inter-marker distance on both chromosomes was taken to be 4 cM and the length of the genetic material outside the boundary markers on each chromosome was 2 cM.

Next, three genes (gene expression measurements) were assigned about each marker, resulting in 300 genes. We assume that the genetic distances from the three gene loci to the marker is so small that any effect between the marker and the genes is of *in cis* nature (this is our link data). Then, for each marker, one gene (the middle one) was randomly assigned a value $\phi_j \in \{0, 1\}$, where $\phi_j = 1$ indicates the presence of an *in cis* effect and for the remaining two genes we assume no *in cis* effect ($\phi_j = 0$). For the middle genes, the probability that $\phi_j = 1$ was taken to be 0.3, which is in line with current estimates (Jansen and Nap, 2004; Morley *et al*, 2004).

To mimic allele-specific expression, the gene expression value $x_{i,j}$ of gene j for individual i was generated from the mixture distribution

$$\begin{cases} N(0,1) & \text{if } \phi_j = 0, \\ \tilde{z}_{i,j,1}N(-1,1) + \tilde{z}_{i,j,2}N(1,1) & \text{if } \phi_j = 1, \end{cases} \tag{7}$$

where $N(a, b)$ is the normal distribution with mean a and variance b, and $\tilde{z}_{i,j,k}$ is the indicator of genotype k for individual i at the marker linked to gene j. Although gene expression values are generated independently from each other, dependence between markers (with *in cis* effects) will imply also some dependence between expression levels.

Phenotype data

The phenotype data were constructed as a linear combination of genetic components, that is, genotype–expression pairs, of six different subtypes (Figure 2). The possibility of single genotype–expression pair simultaneously having more than one active phenotype effect was excluded in the simulation. Thus, we divide the genetic components into three subtypes depending on the mechanism by which they have an effect on the phenotype: genotype effect (G), gene expression effect (E), and genotype × expression effect (GE). Also, we distinguish between marker–gene pairs with and without *in cis* effects. We add an i to denote the presence of *in cis* effect. So, for example, a marker–gene pair of type iE contributes to the phenotype only through the gene expression, although there exists an *in cis* effect on the genotype–expression level.

Based on the above genetic data, a continuous and a binary phenotype were generated. The phenotypes were designed to study which of the above six effect types our model is able to recapture. Also, we wanted to do comparison studies against more traditional models, that is, models based solely on either marker or gene expression data. With this task in mind, a continuous phenotype for individual i was generated as

$$\begin{aligned} y_i = {} & a_1\tilde{z}_{i,s_1,2} + a_2\tilde{z}_{i,s_2,2} + a_3x_{i,s_3} + a_4x_{i,s_4} \\ & + a_5x_{i,s_5}\tilde{z}_{i,s_5,2} + a_6x_{i,s_6}\tilde{z}_{i,s_6,2} + \varepsilon_i \end{aligned}, \qquad (8)$$

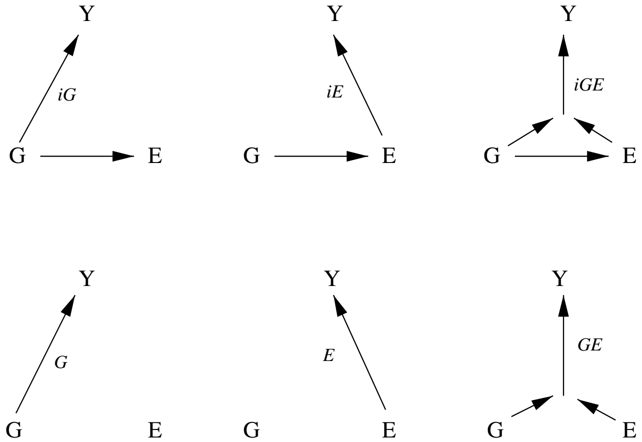

Figure 2 Different genetic effects used in the simulation studies. The mechanism of the effect from the genotypes (G) and the gene expressions (E) to the phenotype (Y) is described using directed graphs. The arrow from G to E indicates genotype-specific gene expressions (*in cis*) and the absence of the arrow means the genotype does not regulate the gene expression. Starting from the left, the graphs indicate the presence of a genotype–phenotype effect, an expression–phenotype effect, and a genotype × expression–phenotype effect, respectively. Also, in each graph, the shortcut notation of the effect type is given using italic fonts.

where the subscripts $s_1, \ldots, s_6$ are randomly chosen indexes of marker–gene pairs of types G, iG, E, iE, GE, and iGE, respectively, and ε_i is normally distributed with mean 0 and variance 1. The factors $a_1, \ldots, a_6$ are the inverses of the sample standard deviations of the six genetic terms, respectively, and their function is to ensure that each term contributes an equal amount of variation to the phenotype. Further, the binary phenotype was defined based on the continuous phenotype as follows:

$$\begin{aligned} & w_i = 1 \quad \text{if } y_i > \frac{1}{n}\sum_{i=1}^{n} y_i \\ & w_i = 0 \quad \text{otherwise} \end{aligned}$$

For realizations of the above phenotype, the heritability, which is a measure of the proportion of the phenotypic variation explained by the genetic components, is typically about 0.6–0.7.

Analyses

We analyzed a realization of the phenotype where the genetic effect components were about equally distributed along the genome and the heritability was 0.69. The proportion of missing values in both the molecular marker data and the gene expression data was taken to be 0.01. These proportions can vary in practice and in addition to reducing the information content in the data, missing values slow down the actual estimation process. The continuous phenotype and the binary phenotype were analyzed separately using combined marker and gene expression data, marker data alone, and gene expression data alone. Because no *trans*-acting variation was included into the simulation, we can easily monitor false positives arising from *in cis* effects in conventional analyses. These six different analyses were implemented using Matlab software on a Pentium IV 2.8 GHz processor. The initial values for the effects θ_j, $j = 1, \ldots, N$, were taken to be zero, and those of the variance terms σ_j^2, $j = 0, \ldots, N$, were initialized to 0.5. The mean value μ was assigned to the sample mean of the phenotypes and the missing values were randomly assigned initial values from their empirical distributions. The Markov chain Monte Carlo (MCMC) algorithm was run for 50 000 rounds ($\approx$2 h) in all simulated examples. In each case, the first 10% of the rounds were considered to be 'burn-in' rounds and were thus discarded from the analysis. Also, to reduce serial correlation, only every 10th round was stored and used in the final summaries. The convergence assessment of the method was made by visually monitoring the chains for several different parameters, mainly the effect coefficients and the error variance. Although the number of the effect coefficients can be very large, in practice one needs to consider only the few chains, as after the burn-in rounds the majority of the effect coefficients are constantly zero. In our simulation studies, the convergence was very fast. In fact, very similar results to those reported are achieved when using samples from the first 20% of the MCMC rounds only.

Results

Continuous trait

Combined data analysis: In Figure 3, the results of the analysis combining marker and gene expression data are

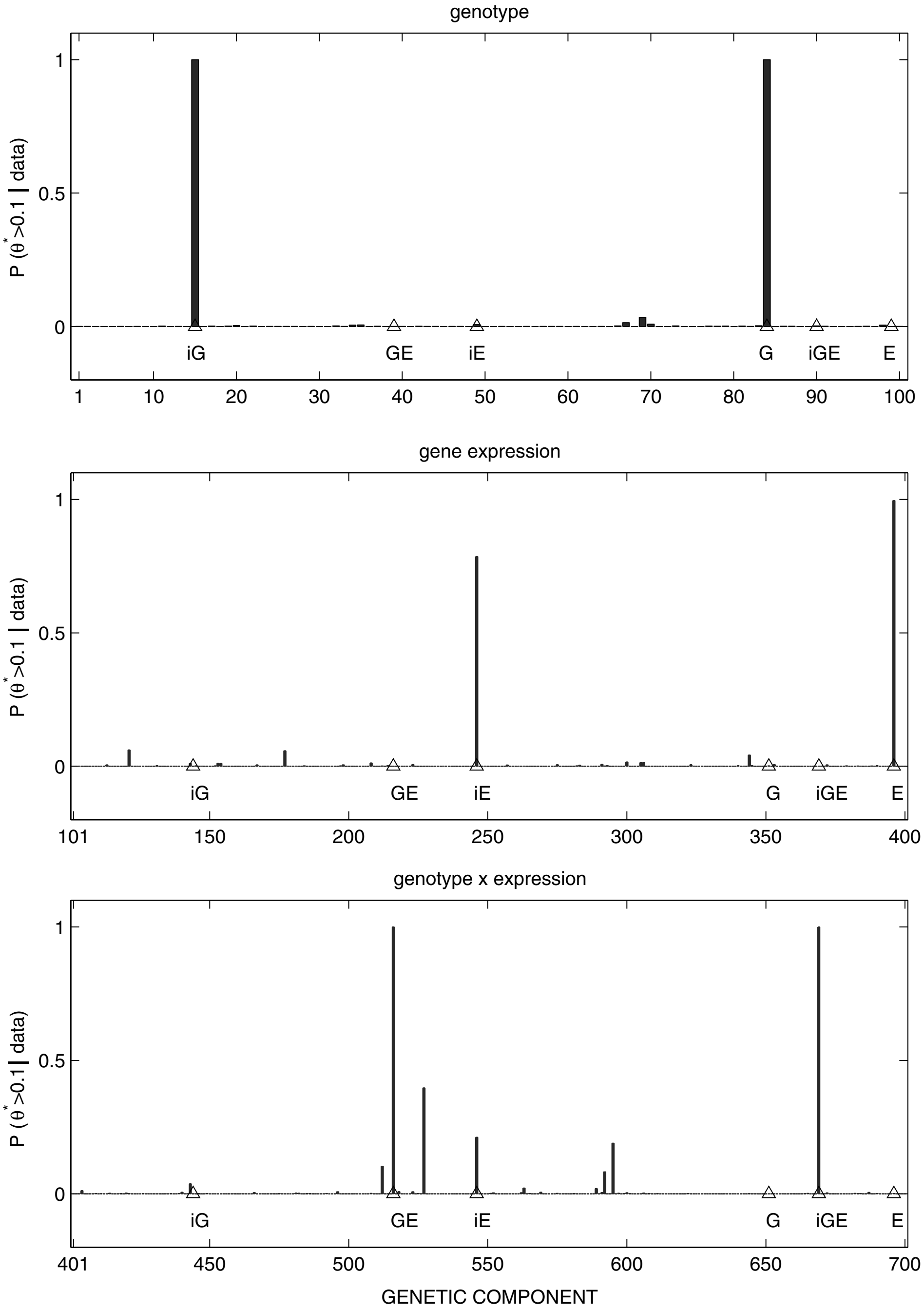

Figure 3 A summary of the combined analysis for the continuous phenotype using both marker and gene expression data. The panels contain the estimated posterior probabilities $P_j^{(0.1)} = P(\theta_j^* > 0.1|\text{data})$ for the genotype effects (top), gene expression effects (middle), and genotype × expression effects (bottom). In all three panels, the positions of the simulated effects are indicated by triangles together with the shortcut notation of the subtype. Note that each panel contains the same genetic section and therefore the corresponding genetic components are vertically leveled.

summarized by the posterior probabilities that the standardized effect size exceeds the given threshold 0.1. The standardized effect (an analog to standardized regression coefficient found in statistical text books) for the genetic component j is given by $\theta_j^* = \theta_j \times \sigma_j / \sigma_p$ where σ_j is the standard deviation of the genetic component j and σ_p is the standard deviation of the phenotype. In the presence of missing values, θ_j^* is calculated on every MCMC round using σ_j of the imputed data. Further in the binary case, σ_p is the standard deviation of the liability phenotype, which changes every MCMC round. The above posterior probability for the genetic component j can be written as $P_j^{(0.1)} = P(\theta_j^* > 0.1 \mid \text{data})$, where 'data' refers to the observed genetic data and the phenotype. In summary of Figure 3, altogether nine genetic components had elevated posterior probabilities $P_j^{(0.1)}$, where this probability was distinctive (mostly 1.0) for all six simulated components, and it was equal or less than 0.4 for the others. The number of nonzero effects is controlled adaptively in the analysis (their number depends on the data at hand).

To summarize the number of influential components, $I_c^{(0.1)} = \sum_j 1(\theta_j^* > 0.1)$, in the genetic model, Table 1 presents the posterior probabilities for different numbers of components, whose standardized effect size simultaneously exceeds the threshold 0.1, that is, $P(I_c^{(0.1)} = n \mid \text{data})$. Note that the distribution only supports numbers in the range [6, 10] and that the support is clearly highest at the correct number six. In Table 2, we have calculated the conditional probability $Q_{j,k}^{(0.1)} = P(\theta_j^* > 0.1 | \theta_k^* > 0.1, \text{data})$ for all pairs {*j*, *k*}, formed by genetic components whose probability $P_j^{(0.1)}$ exceeds 0.1. $Q_{j,k}^{(T)}$ is the posterior probability that the standardized effect of the genetic component *j* exceeds the threshold *T* conditional that the standardized effect of the genetic component *k* exceeds the same threshold. These *Q*-summaries allow the detection of alternating components in the genetic model. From Table 2 and by visually studying the MCMC paths of the standardized effects (Figure 4), it can be seen that the expression effect ($j = 246$) and the genotype × expression effect ($j = 546$), associated to the genetic component of type *iE*, are complementary, and thus only one of the two at a time contributes a nonzero effect into the genetic model. The threshold value $T = 0.1$ was chosen subjectively. Experiments with different threshold values indicated that the above summary statistics are robust to the chosen value. The use of a smaller threshold value $T < 0.1$, although increasing the number of components with nonzero $P_j^{(T)}$ values, rarely had an effect on the number of components with higher $P_j^{(T)}$ values, say greater than 0.1. Also, the effect on the distribution of the number of influential components was negligible.

Sole marker analysis: In the top-left panel of Figure 5, the standardized genotype effects are summarized by the component probabilities $P_j^{(0.1)}$. From the results, we can locate four markers (genetic components), which all satisfy the condition $P_j^{(0.1)} > 0.1$. Further, by studying the *Q*-summaries, the pairwise probabilities $Q_{j,k}^{(0.1)}$ (table not shown), it can be concluded that the genetic components 83 and 86 are complementary. Thus, the results suggest that we are able to locate three putative markers only, two correct ones having main genotype effects (types *iG* and *G*) and a 'false-positive' one ($j = 45$) located between the type *GE* component and the type *iE* component. These conclusions are further supported by Table 1, where the highest probability 0.60 is assigned to the case $n = 3$.

Sole gene expression analysis: In the lower-left panel of Figure 5, the standardized gene expression effects are

Table 1 The distributions of the number of influential components

Analysis	n										
	0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Continuous trait											
Combined data	0.00	0.00	0.00	0.00	0.00	0.00	**0.50**	**0.31**	**0.13**	0.04	0.01
Marker data	0.00	0.00	**0.31**	**0.60**	0.08	0.01	0.00	0.00	0.00	0.00	0.00
Expression data	0.00	0.00	0.00	0.01	**0.41**	**0.38**	**0.15**	0.05	0.01	0.00	0.00
Binary trait											
Combined data	0.00	0.00	0.00	0.02	**0.43**	**0.38**	**0.13**	0.03	0.00	0.00	0.00
Marker data	0.00	0.01	**0.77**	**0.20**	0.02	0.00	0.00	0.00	0.00	0.00	0.00
Expression data	0.00	0.00	**0.40**	**0.37**	**0.18**	0.04	0.01	0.00	0.00	0.00	0.00

The posterior probabilities for different numbers of influential components $P(I_c^{(0.1)} = n \mid \text{data})$, where *n* is given in the top row and the remaining rows correspond to the six different analyses given in the first column. For clarification, the posterior probabilities exceeding 0.1 are highlighted in bold.

Table 2 Pairwise conditional summaries

j	$P_j^{(0.1)}$	k								
		15	*84*	*246*	*396*	*516*	*527*	*546*	*595*	*669*
15	1.00		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
84	1.00	1.00		1.00	1.00	1.00	1.00	1.00	1.00	1.00
246	0.79	0.79	0.79		0.79	0.79	0.57	**0.01**	0.75	0.79
396	1.00	1.00	1.00	1.00		1.00	0.99	1.00	1.00	1.00
516	1.00	1.00	1.00	1.00	1.00		1.00	1.00	1.00	1.00
527	0.40	0.40	0.40	0.29	0.40	0.40		0.80	0.45	0.40
546	0.21	0.21	0.21	**0.00**	0.21	0.21	0.43		0.25	0.21
595	0.19	0.19	0.19	0.18	0.19	0.19	0.22	0.23		0.19
669	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	

The *Q*-summaries of the analysis of the continuous phenotype utilizing combined marker and gene-expression data. The first column contains the indexes *j* of the genetic components whose posterior probability $P_j^{(0.1)} = P(\theta_j^* > 0.1 \mid \text{data})$ exceeds 0.1 (see Figure 3). The second column contains the actual values of those probabilities $P_j^{(0.1)}$. In the remaining submatrix we give the pairwise conditional probabilities $Q_{j,k}^{(0.1)} = P(\theta_j^* > 0.1 \mid \theta_k^* > 0.1, \text{data})$, where *k* is given in the top row. Small *Q*-values, indicating possible complementary components are highlighted in bold.

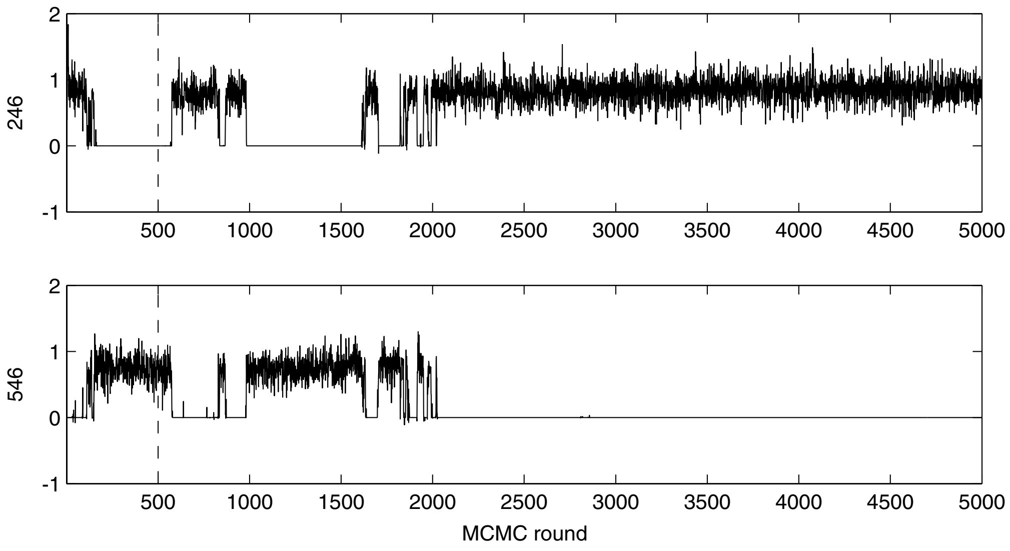

Figure 4 The MCMC sample paths of the standardized effects of the genetic components 246 and 546. The vertical dashed line indicates the end of the burn-in period.

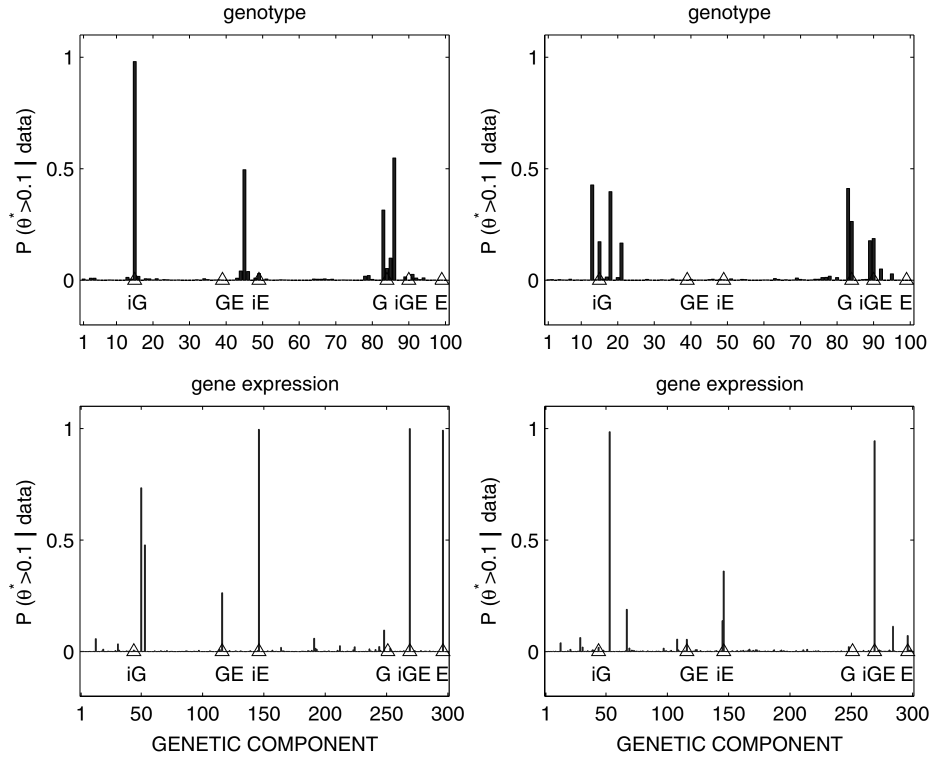

Figure 5 A summary of the analysis for both the continuous phenotype (left panels) and the binary phenotype (right panels) using only marker data (upper panels) and only gene expression data (lower panels). The notation in the panels is as in Figure 3.

summarized by the component probabilities $P_j^{(0.1)}$. We were able to locate all the genes (genetic components) that contribute to the continuous phenotype through their expression (types *E*, *iE*, *GE*, and *iGE*). Also, there is some evidence ($P_j^{(0.1)}>0.1$) about the two simulated components with no expression effects: two components ($j=50$ and $j=53$) close to the component of type *G* and one ($j=248$) at the component of type *iG*. This can be explained by the correlation between the gene expression values and the phenotype, which is induced by the high correlation between close markers and their *in cis* effects (recall that about 30% of the markers have *in cis* effects). By studying the MCMC sample paths of components 50 and 53 in Figure 6, the strong interaction between their effect sizes becomes apparent; if either of the components obtains a zero

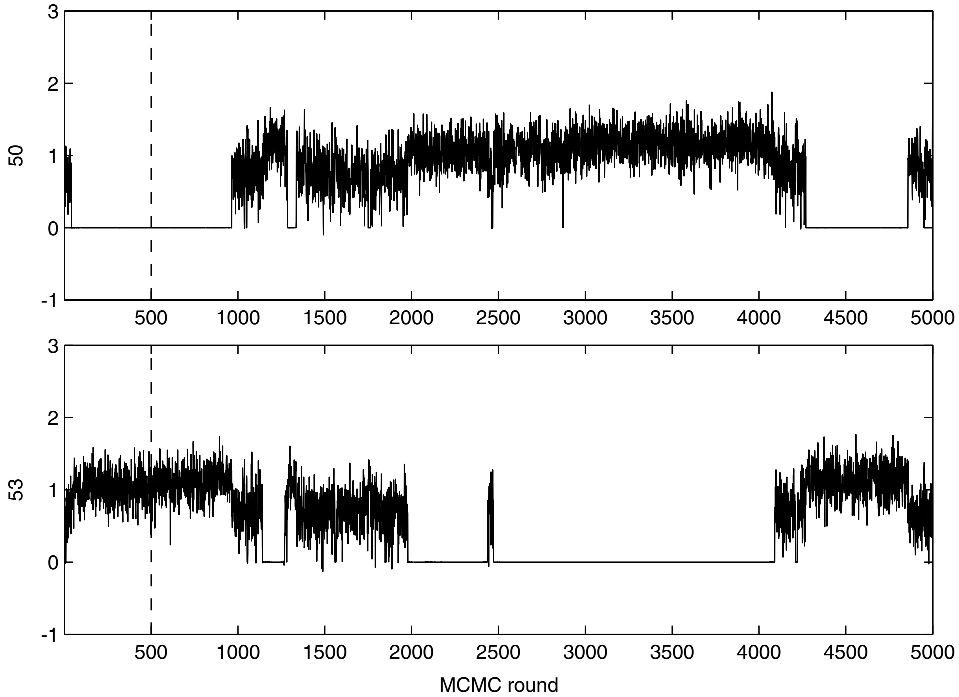

Figure 6 The MCMC paths of the standardized effects of the genetic components 50 and 53 from the analysis of the continuous phenotype using gene expression data only. The vertical dashed line indicates the end of the burn-in period. Note that if one of the paths vanishes, the other compensates it by taking on a higher value.

effect, the other compensates its absence by taking on higher values. Thus, again we can make the conclusion that both effects attempt to model the same genetic effect (type *G*). This conclusion cannot be made from the *Q*-summaries, as $Q^{(0.1)}_{50,53}=0.50$ and $Q^{(0.1)}_{53,50}=0.33$. In summary, we were able to locate six putative genes, with one ($j=248$) having a posterior probability value of 0.1 only. This result is also supported by Table 1, where although some evidence is assigned to the case $n=6$, the highest probabilities are obtained for the cases $n=4$ and $n=5$.

Binary trait

Combined data analysis: In Figure 7, the results of the analysis, utilizing both marker data and gene expression data, are summarized by the probabilities $P^{(0.1)}_j$. Eight genetic components obtained $P^{(0.1)}_j$ values higher than 0.1, although the $P(I^{(0.1)}_c=n\,|\,\text{data})$ values in Table 1 supported strongly the existence of only four or five influential components. From the *Q*-summaries (Table 3), it is apparent that components 15 and 16 as well as 84 and 85 are complementary and represent alternative signals from the same underlying component. Thus, the model suggests the existence of six genetic components from which the two with the smallest $P^{(0.1)}_j$ values are false positives. Further candidates can be unveiled by considering genetic components with smaller $P^{(0.1)}_j$ values than 0.1. However, their existence is not supported by the $P(I^{(0.1)}_c=n\,|\,\text{data})$ values. Not surprisingly, the performance of the method using a binary phenotype is slightly worse than that of the continuous counterpart, which can be explained by the loss of information in the dichotomization process.

Sole marker analysis: In the analysis of the binary phenotype, several high $P^{(0.1)}_j$ values showed up in the vicinity of the two markers with main genotype effects (types *iG* and *G*) (see the upper-right panel of Figure 5). Also high values were found around the component of type *iGE*. The $P(I^{(0.1)}_c=n\,|\,\text{data})$ values in Table 1 and the *Q*-summaries (table not shown) support the existence of two or three putative markers.

Sole gene expression analysis: In this analysis of the binary phenotype, six genetic components (gene) were assigned $P^{(0.1)}_j$ values greater than 0.1, with two having values close to 1 (components 53 and 269). Here component 269 is a real simulated expression effect (type *iGE*) and component 53 is an artifact arising from the nearby simulated effect of type *iG*. Component 146 (type *iE*) was assigned a $P^{(0.1)}_j$ value 0.39 and the remaining three putative components had values about 0.2 or less (see the lower-right panel of Figure 5). The $P(I^{(0.1)}_c=n\,|\,\text{data})$ values in Table 1 supported the existence of two to four putative genes.

Heritability and effect estimation

In addition to producing location estimates of putative genetic components, the analysis provides posterior heritability and effect estimates. If the model is able to successfully capture the phenotypic variation due to the genetic components, then accurate heritability estimates are obtained from the formula $(1/r)\sum_{t=1}^{r}(\sigma_p^{2(t)}-\sigma_0^{2(t)})/\sigma_0^{2(t)}$, where $\sigma_p^{2(t)}$ is the phenotypic variance, $\sigma_0^{2(t)}$ is the error variance at round t, and r is the total number of MCMC rounds. In the continuous case, the phenotypic variance is constant over the MCMC rounds, and in the binary case, it is calculated using the

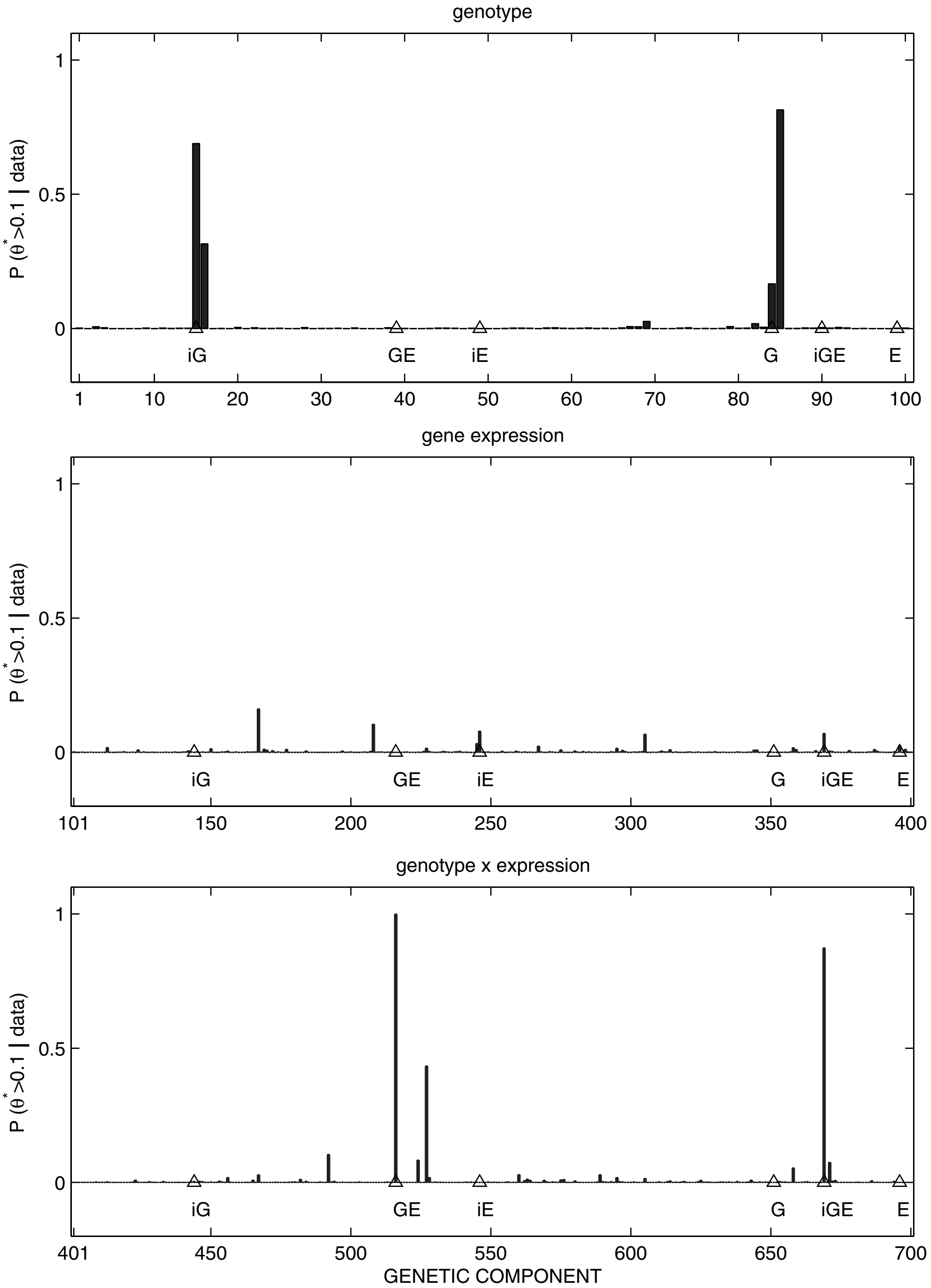

Figure 7 A summary of the combined analysis for the binary phenotype using both marker and gene expression data. The panels and the notation are as in Figure 3.

liability phenotype. Note also that in the binary case, the error variance is 1 by definition. The posterior heritability estimates based on all six analyses are given in Table 4. We emphasize the huge drop in the accuracy of the estimates, when only a subset of the original genetic components used to simulate the phenotype is included as data in the analysis. Therefore, some caution should be taken when reporting or making conclusions on heritability estimates based on analyses of real data sets.

The accuracy of the effect estimates provided by the model (results not shown) is similar to that of the heritability estimates. For example, the posterior distributions of standardized genetic effects with $P_j^{(0.1)}$ values close to 1, in the combined analysis of the continuous phenotype, are all highly concentrated around 1, which is the correct value of the standardized simulated effects. Figures 4 and 6 are typical examples of the MCMC paths of standardized effects for influential genetic components, whose $P_j^{(0.1)}$ value is smaller than 1. As indicated in the figures, their distributions have multiple modes, which need to be taken into account when conclusions are drawn.

Table 3 Pairwise conditional summaries

j	$P_j^{(0.1)}$	k							
		15	*16*	*84*	*85*	*167*	*516*	*527*	*669*
15	0.69		**0.01**	0.80	0.66	0.53	0.69	0.93	0.68
16	0.32	**0.01**		0.21	0.34	0.48	0.32	0.08	0.32
84	0.17	0.19	0.11		**0.00**	0.06	0.17	0.25	0.18
85	0.81	0.79	0.88	**0.01**		0.93	0.81	0.72	0.80
167	0.16	0.12	0.25	0.06	0.19		0.16	0.04	0.17
516	1.00	1.00	1.00	1.00	1.00	1.00		1.00	1.00
527	0.43	0.58	0.11	0.66	0.38	0.11	0.43		0.42
669	0.87	0.86	0.90	0.93	0.86	0.93	0.87	0.84	

The *Q*-summaries of the analysis of the binary phenotype utilizing both the marker and the gene-expression data. Notations as in Table 2.

Table 4 Heritability estimates

Analysis	*Median*	*95% interval*
Continuous trait		
Combined data	0.67	[0.58, 0.74]
Marker data	0.35	[0.19, 0.47]
Expression data	0.42	[0.27, 0.53]
Binary trait		
Combined data	0.62	[0.46, 0.71]
Marker data	0.35	[0.16, 0.51]
Expression data	0.35	[0.13, 0.52]

A summary of posterior heritability estimates calculated from the MCMC samples. The 95% credible interval is calculated so that 2.5% of the samples from both ends of the distribution are left out. The heritability, calculated from the simulated data, using the continuous phenotype and the true error variance, is 0.69.

Real data example

Additionally, in order to experiment on real data, we applied our method to the publicly available data set of Brem *et al* (2002). This *Saccharomyces cerevisiae* data consist of gene expression measurements and marker genotypes in 40 haplotypes (segregants) from a cross between a laboratory and a wild strain of Yeast. As a starting point of our analysis, we chose 102 genes that were reported by Brem *et al* (2002) to belong to one of the first five gene groups, which contain the known members of different pathways. Based on the 570 expression QTLs (gene–marker pairs) reported by Brem *et al* (2002), we identified 12 such markers to whom at least one of our chosen 102 genes was linked. To perform genetical genomics analysis, we treated one of the genes, namely gene YLR089C, as the expression phenotype and explained its variation by using the remaining 101 genes, 12 markers, and 101 marker–gene pairs as covariates in model (1). Note that noninformative prior (omitting inter-marker distances) was assumed for missing genotype data.

We ran four separate MCMC chains (runs I–IV) each of length 100 000 ($\approx$1 h) using different starting values. First 50 000 rounds from each chain were discarded because of 'burn-in' and every 50th sampled values were stored and utilized in actual MCMC estimation. The convergence of each chain was checked by visually inspecting MCMC paths for several different parameters. Although all chains seemed to be converged well, we found some variation between the results. Such a behavior is evidently owing to the high correlation between genes within the same pathway and the fact that the sample size (number of individuals) was actually quite small.

In three (runs I–III) out of the four cases, we located a significant effect at component 44 (gene YMR108W), with probability $P_{44}^{(0,1)} = 0.69$ in run I and with probability $P_{44}^{(0,1)} = 0.98$ in runs II and III. Based on the *Q*-summaries (results not shown), the small probability value 0.69 at component 44 in run I can be explained by the existence of its complementary effect with the component 49, having probability $P_{49}^{(0,1)} = 0.33$. In the remaining case (run IV), another gene located in the same pathway was identified, namely the component 42 (gene YHR208W), with probability $P_{42}^{(0,1)} = 1.0$. Both identified genes are in the same pathway with the expression phenotype. In addition to the main terms, a genotype–expression interaction term arose at gene YMR108W in run I and at gene YER073W in runs II and III. Not surprisingly, gene YER073W is also located in the same pathway as the expression phenotype (gene) and all the interaction terms are linked to the same marker, '2435_at_ × 00'.

This example illustrates difficulties that can occur in applications of the method to real life data sets, that is, the method seems to suffer convergence problems owing to the small sample size (no. of individuals) and the highly correlated candidates (and perhaps noisy phenotype). Note that Wang *et al* (2005) also recognized potential mixing problems of their method in the presence of highly correlated genetic components (closely linked markers) and small sample size. To alleviate these problems in extremely small data sets like here, we suggest that future studies should run several different MCMC chains and base their estimates on the pooled MCMC samples, where samples from several different MCMC chains are combined together.

Discussion

Effect components

We have presented a novel Bayesian sparse method, which allows us to simultaneously utilize measurements from multiple data sources (molecular markers and gene expression microarrays) to model phenotype. The benefit of the method compared to conventional phenotype–marker or phenotype–expression association method is the possibility to consider genotype × expression interactions by introducing marker–gene pairs as link data.

Also, available environmental (nongenetic) covariates of discrete or continuous type can be included into the model by coding them in a similar manner as 'markers' with multiple variants or 'gene expression' measurements, respectively. Protein-expression measurements (Sellers and Yates, 2003) can be included as 'gene expressions' in a similar fashion. In the presence of readily available database information (eg GO, KEGG; see Ashburner *et al*, 2000) about gene × gene and expression × expression interactions involved in known pathways (Thomas, 2005), the model can be easily expanded to include known pairs of epistatic markers (Conti *et al*, 2003) and known expression × expression interaction determinants. Also genetic × nongenetic interactions can be considered. If the sample size is large, one can even search through all possible pairwise combinations by incorporating epistatic effects into the oversaturated model following Zhang and Xu (2005).

Resolution

The mapping resolution can be increased by introducing new marker points, the so-called pseudo-markers, along a discrete grid (Sen and Churchill, 2001) or by adding a random QTL position into each marker interval (Wang *et al*, 2005). The genotype information at these new markers is handled as missing values and their patterns are predicted based on the genetic distances and the observed genetic configuration of the surrounding markers. In addition to marker data, one could also use *cis-* and *trans-*acting gene expression information to impute/predict pseudo-marker genotypes. Note that one can still include the original gene expression measurements as putative candidates into model (1). However, these treatments as such are applicable only for controlled crosses. With small data sets, introducing pseudo-markers on small intervals may enrich correlateness (colinearity) between candidates, which again may potentially lead to mixing problems of the sampler.

Alternative model considered: An intuitive picture of the mechanism leading to sparseness in the Bayesian model is provided by the following reasoning (see Figueiredo, 2003). Replace Jeffreys' prior $p(\sigma_j^2) \propto 1/\sigma_j^2$ by the double exponential prior $p(\sigma_j^2) = (\gamma/2)\exp(-(\gamma/2)\sigma_j^2)$. It is then possible to analytically integrate out σ_j^2 from the prior distribution of the effect θ_j resulting in the well-known Laplacian distribution $p(\theta_j|\gamma) = (\sqrt{\gamma}/2)\exp(-\sqrt{\gamma}|\theta_j|)$. Thus, replacing Jeffreys' prior with the double exponential prior is identical to assuming *a priori* that the effects θ_j have a Laplacian (sparse) distribution with a common parameter γ. A drawback of such an approach is that the Laplacian prior does not allow the application of Gibbs sampling steps in updating the effects, and also tuning of the extra parameter γ can be difficult and time consuming. We performed various experiments using the Laplacian prior with different Metropolis–Hastings steps, but we finally abandoned the approach owing to serious mixing and convergence problems. However, if these problems are overcome in the future, Laplacian prior provides the user a way to control the number of influential components (nonzero effects) or the degree of sparseness in the model, where lack of such parameter can be thought of as a drawback of the present adaptive approach. Note that there are various other approaches that allow user to control the amount of sparseness in the data (eg Tibshirani, 1996; Meuwissen *et al*, 2001; Sillanpää and Bhattacharjee, 2005; Zhang *et al*, 2005).

Human data sets: The method was presented for controlled line crossing experiments of inbred animals or plants with two genotype combinations. However, the method can be easily generalized for human data sets (population-based samples) using single nucleotide polymorphism (SNP) markers, where there are only three genotype combinations (two coefficients are required in the model with their own variance components). Difference between using this method for analyzing inbred line cross F2 and human population-based samples of SNPs is that in F2 one can use information from other markers in the prior for missing marker imputation. For microsatellite markers having multiple genotypes at single locus, one can alternatively assume exchangeability and fit common variance for all the effects at single location (see Meuwissen *et al*, 2001). Additional complexity in human studies that is expected to occur in simultaneous marker and expression analysis is the population stratification – confounding owing to unobserved population substructure. This problem has been considered in marker association studies (Lander and Schork, 1994; Sillanpää and Bhattacharjee, 2005) and in expression association studies (Gibson, 2003; Kraft and Horvath, 2003; Kraft *et al*, 2003; Lu *et al*, 2004).

Comparing results between experiments: The use of the standardized effects enables direct comparisons of the effect sizes of genetic components with different variances and of experiments with different phenotypic variances. An alternative approach would be to normalize the genetic components and the phenotype in advance. However, in the binary case and when there is missing data present, the normalization needs to be performed on every MCMC round. In studies utilizing data from a single source, this problem is seldom present, as the components have naturally somewhat similar variances. For example, the component variance of marker data from a backcross study is about 0.25, and gene expression measurements are of equal scale after preprocessing. Further, the use of standardized effects makes the comparison of separate studies more easy and informative. This is important especially now when combining data into meta-analysis has become popular (see Sillanpää and Auranen, 2004). In our study, the use of Jeffreys' scale invariant prior allows us to perform the actual analyses without normalizing the genetic components in advance. However, if some other prior is used, the scale and thus the normalization of the data can prove to be extremely important.

A Matlab implementation of the method is available from the authors upon request.

Acknowledgements

We are grateful to Miguel Perez-Enciso for discussions. This work was supported by research grant (202324) from the Academy of Finland.

References

Albert JH, Chib S (1993). Bayesian analysis of binary and polychotomous response data. *J Am Stat Assoc* **88**: 669–679.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, More M (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**: 25–29.

Auger DL, Gray AD, Ream TS, Kato A, Coe Jr EH, Birchler JA (2005). Nonadditive gene expression in diploid and triploid hybrids of maize. *Genetics* **169**: 389–397.

Aune TM, Parker JS, Mass K, Liu Z, Olson NJ, Moore JH (2004). Co-localization of differentially expressed genes and shared susceptibility loci in human autoimmunity. *Genet Epidemiol* **27**: 162–172.

Basten CJ, Weir BS, Zeng Z-B (1994). Zmap – a QTL Cartographer. In: Smith C, Gavora JS, Benkel B, Chesnais J, Fairfull W, Gibson JP, Kennedy BW, Burnside EB (eds) *Proceedings of the 5th World Congress on Genetics Applied to Livestock Production: Computing Strategies and Software*. Vol 22, Organizing Committee, 5th World Congress on Genetics Applied to Livestock Production: Guelph, Ontario, Canada. pp 65–66.

Basten CJ, Weir BS, Zeng Z-B (2003). *QTL Cartographer, Version 117*. Department of Statistics, North Carolina State University: Raleigh, NC.

Brem R, Yvert G, Clinton R, Kruglyak L (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**: 752–755.

Broman KW, Speed TP (2002). A model selection approach for identification of quantitative trait loci in experimental crosses. *J R Stat Soc B* **64**: 641–656.

Butte A (2002). The use and analysis of microarray data. *Nat Rev Drug Discov* **1**: 951–958.

Bystrykh L, Weersing E, Dontje B, Sutton S, Pletcher MT, Wiltshire T *et al* (2005). Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat Genet* **37**: 225–232.

Casella G, George EI (1992). Explaining the Gibbs sampler. *Am Stat* **46**: 164–174.

Chesler EJ, Lu L, Shou SM, Qu YH, Gu J, Wang JT *et al* (2005). Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat Genet* **37**: 233–242.

Chib S, Greenberg E (1995). Understanding the Metropolis–Hastings algorithm. *Am Stat* **49**: 327–335.

Conti DV, Cortessis V, Molitor J, Thomas DC (2003). Bayesian modeling of complex metabolic pathways. *Hum Hered* **56**: 83–93.

Darvasi A (2003). Gene expression meets genetics. *Nature* **422**: 269–270.

Devlin B, Roeder K, Wasserman L (2003). Analysis of multilocus models of association. *Genet Epidemiol* **25**: 36–47.

Doerge RW (2002). Mapping and analysis of quantitative trait loci in experimental populations. *Nat Rev Genet* **3**: 43–52.

Figueiredo MAT (2003). Adaptive sparseness for supervised learning. *IEEE Trans Pattern Anal Mach Intell* **25**: 1150–1159.

Geman S, Geman D (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intell* **6**: 721–741.

Gibson G (2003). Population genomics: celebrating individual expression. *Heredity* **90**: 1–5.

Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC (2004). A global test for group of genes: testing association with a clinical outcome. *Bioinformatics* **20**: 93–99.

Hastings WK (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**: 97–109.

Heighway J, Bowers NL, Smith S, Betticher DC, Santibanez Koref F (2005). The use of allelic expression differences to ascertain functional polymorphisms acting *in cis*: analysis of MMP1 transcripts in normal lung tissue. *Ann Hum Genet* **69**: 127–133.

Hubner N, Wallace CA, Zimdahl H, Petretto E, Schulz H, Maciver F *et al* (2005). Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat Genet* **37**: 243–253.

Jansen RC (2003). Studying complex biological systems using multifactorial perturbation. *Nat Rev Genet* **4**: 145–151.

Jansen RC, Nap J-P (2001). Genetical genomics: the added value from segregation. *Trends Genet* **17**: 388–391.

Jansen RC, Nap J-P (2004). Regulating gene expression: surprises still in store. *Trends Genet* **20**: 223–225.

Jiang C, Zeng Z-B (1997). Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. *Genetica* **101**: 47–58.

Kell DB (2002). Genotype–phenotype mapping: genes as computer programs. *Trends Genet* **18**: 555–559.

Kilpikari R, Sillanpää MJ (2003). Bayesian analysis of multilocus association in quantitative and qualitative traits. *Genet Epidemiol* **25**: 122–135.

Knight JC (2004). Allele-specific gene expression uncovered. *Trends Genet* **20**: 113–116.

Kopp A, Graze RM, Xu S, Carroll SB, Nuzhdin SV (2003). Quantitative trait loci responsible for variation in sexually dimorphic traits in *Drosophila melanogaster*. *Genetics* **163**: 771–787.

Kraft P, Horvath S (2003). The genetics of gene expression and gene mapping. *Trends Biotechnol* **21**: 377–378.

Kraft P, Schadt E, Aten J, Horvath S (2003). A family-based test for correlation between gene expression and trait values. *Am J Hum Genet* **72**: 1323–1330.

Lan H, Stoehr JP, Nadler ST, Schueler KL, Yandell BS, Attie AD (2004). Dimension reduction for mapping mRNA abundance as quantitative traits. *Genetics* **164**: 1607–1614.

Lander ES, Schork NJ (1994). Genetic dissection of complex traits. *Science* **265**: 2037–2048.

Lo HS, Wang Z, Hu Y, Yang HH, Gere S, Buetow KH *et al* (2003). Allelic variation in gene expression is common in the human genome. *Genome Res* **13**: 1855–1862.

Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirchorn JN (2003). Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common diseases. *Nat Genet* **33**: 177–182.

Lopes H, West M (2004). Bayesian model assessment in factor analysis. *Stat Sinica* **14**: 41–67.

Lu Y, Liu P-Y, Liu Y-J, Xu F-H, Deng H-W (2004). Quantifying the relationship between gene expression and trait values in general pedigrees. *Genetics* **168**: 2395–2405.

Meuwissen THE, Hayes BJ, Goddard ME (2001). Prediction of total genetic value using genome-wide dense marker map. *Genetics* **157**: 1819–1829.

Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS *et al* (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**: 743–747.

Perez-Enciso M (2004). *In silico* study of transcriptome genetic variation in outbred populations. *Genetics* **166**: 547–554.

Perez-Enciso M, Toro MA, Tenenhaus M, Gianola D (2003). Combining gene expression and molecular marker information for mapping complex trait genes: a simulation study. *Genetics* **164**: 1597–1606.

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

Risch N, Merikangas K (1996). The future of genetic studies of complex human diseases. *Science* **273**: 1616–1617.

Ronald J, Akey JM, Whittle J, Smith EN, Yvert G, Kruglyak L (2005). Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays. *Genome Res* **15**: 284–291.

Schadt EE, Monks S, Drake T, Lusis A, Che N, Colinayo V *et al* (2003). The genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**: 297–302.

Sellers TA, Yates JR (2003). Review of proteomics with applications to genetic epidemiology. *Genet Epidemiol* **24**: 83–98.

Sen S, Churchill GA (2001). A statistical framework for quantitative trait mapping. *Genetics* **159**: 371–387.

Shevade SK, Keerthi SS (2003). A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics* **19**: 2246–2253.

Sillanpää MJ, Arjas E (1998). Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics* **148**: 1373–1388.

Sillanpää MJ, Auranen K (2004). Replication in genetic studies of complex traits. *Ann Hum Genet* **68**: 646–657.

Sillanpää MJ, Bhattacharjee M (2005). Bayesian association-based fine mapping in small chromosomal segments. *Genetics* **169**: 427–439.

Sillanpää MJ, Corander J (2002). Model choice in gene mapping: what and why. *Trends Genet* **18**: 301–307.

The International HapMap Consortium (2003). The International HapMap project. *Nature* **426**: 789–796.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature* **437**: 1299–1320.

Thomas DC (2005). The need for a systematic approach to complex pathways in molecular epidemiology. *Cancer Epidemiol Biomarkers Prev* **14**: 557–559.

Tibshirani R (1996). Regression shrinkage and selection via the lasso. *J R Stat Soc B* **58**: 267–288.

Wang H, Zhang Y-M, Li X, Masinde GL, Mohan S, Baylink DJ *et al* (2005). Bayesian shrinkage estimation of QTL parameters. *Genetics* **170**: 465–480.

Watts JA, Morley M, Burdick JT, Fiori JL, Ewens WJ, Spielman RS *et al* (2002). Gene expression phenotype in heterozygous carriers of Ataxia Telangiectasia. *Am J Hum Genet* **71**: 791–800.

Wayne ML, McIntyre LM (2002). Combining mapping and arraying: an approach to candidate gene identification. *Proc Natl Acad Sci USA* **99**: 14903–14906.

West M (2003). Bayesian factor regression models in the 'large p, small n' paradigm. In: Bernando JM, Bayarri MJ, Berger JO, Dawid AP, Heckerman D, Smith AFM, West M (eds) *Bayesian Statistics 7*. Oxford University Press: Oxford. pp 723–732.

Xu S (2003). Estimating polygenic effects using markers of the entire genome. *Genetics* **163**: 789–801.

Yan H, Yuan W, Velculescu VE, Vogelstein B, Kinzler KW (2002). Allelic variation in human gene expression. *Science* **297**: 1143.

Zhang M, Montooth KL, Wells MT, Clark AG, Zhang D (2005). Mapping multiple quantitative trait loci by Bayesian classification. *Genetics* **169**: 2305–2318.

Zhang Y-M, Xu S (2005). A penalized maximum likelihood method for estimating epistatic effects of QTL. *Heridity* **95**: 96–104.

Appendix A1

Quantitative phenotype and no missing data

By assuming no missing values and as the prior distributions of the unknown variables θ are of conjugate form, we can apply the Gibbs sampling algorithm (Geman and Geman, 1984; Casella and George, 1992). In the following sketch of the Gibbs sampling algorithm, the exact fully conditional marginal distributions needed to draw dependent samples from the unknown joint posterior distribution of μ, θ, and σ^2 are given in an explicit form.

The single-site Gibbs sampling algorithm:

First initialize μ, θ, and σ^2: $\mu^{(0)} = (1/n)\sum_{i=1}^{n} y_i$, $\theta_j^{(0)} = 0$, for $j = 1, \ldots, N$ and $\sigma_j^{2(0)} = 0.5$, for $j = 0, \ldots, N$. The superscript denotes the ordering of the samples, starting from the initial values (0) up to the total number of sampling rounds. The asterisk (*) in the superscript refers to the most recent value, that is, at round t it corresponds to either t or $t-1$ depending on whether the parameter has been updated on this round or not, respectively. Next, repeat the following steps 1–4 until the prespecified number of samples (sampling rounds) of $\mu^{(t)}$, $\theta^{(t)}$, and $\sigma^{2(t)}$ are produced. Note that the generating distributions are the fully conditional marginal distributions of the random variable being updated.

1. Generate the sample $\mu^{(t)}$ from a normal distribution with mean
$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{N}\theta_j X_{i,j}\right)$$
and variance $(1/n)\sigma_0^{2(t-1)}$.
2. For $j = 1, \ldots, N$, one at a time, generate a sample $\theta_j^{(t)}$ from a normal distribution with mean
$$\left(\sum_{i=1}^{n} X_{i,j}^2 + \frac{\sigma_0^{2(*)}}{\sigma_j^{2(*)}}\right)^{-1} \sum_{i=1}^{n} X_{i,j}\left(y_i - \mu^{(*)} - \sum_{k\neq j}\theta_k^{(*)} X_{i,k}\right)$$
and variance
$$\left(\sum_{i=1}^{n} X_{i,j}^2 + \frac{\sigma_0^{2(*)}}{\sigma_j^{2(*)}}\right)^{-1} \sigma_0^{2(*)}$$
3. Generate $\sigma_0^{2(t)}$ from
$$\frac{1}{s_n}\sum_{i=1}^{n}\left(y_i - \mu^{(*)} - \sum_{j=1}^{N}\theta_j^{(*)} X_{i,j}\right)^2$$
where s_n is a random variable with a χ^2 distribution on n degrees of freedom.
4. For $j = 1, \ldots, N$, generate new samples $\sigma_j^{2(t)}$ from $\theta_j^{2(t)}/s_1$.

Binary phenotype

The use of the above Gibbs algorithm as such requires a quantitative phenotype. However, with only a few modifications, the same algorithm can be used in the binary phenotype case. For a binary variable w_i, the goal is to model the probability $P(w_i = 1)$, more precisely $P(w_i = 1 \mid \theta, \sigma^2, D)$. First, let us assume a latent continuous variable y_i such that

$$w_i = \begin{cases} 1 & \text{if and only if } y_i > 0 \\ 0 & \text{otherwise} \end{cases} \tag{A.1}$$

where y_i is given by equation (4) with error term $\varepsilon_i \sim N(0,1)$. It then follows that $P(w_i = 1) = P(y > 0 \mid \theta, \sigma^2, D) = \Phi(\mu + \sum_{j=1}^{N}\theta_j X_{i,j})$, where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal. Thus, with the above assumptions, the binary phenotype is modeled via the probit link function (a Bayesian probit model).

Applying the Bayesian probit model requires the use of data augmentation (Albert and Chib, 1993; Kilpikari and Sillanpää, 2003), that is, sampling continuous latent variables y_i at each round. This is carried out by sampling y_i directly from the posterior distributions

$$N_-\left(\mu+\sum_{j=1}^{N}\theta_j X_{i,j},1\right) \text{ if } w_i=0 \qquad \text{(A.2)}$$

$$N_+\left(\mu+\sum_{j=1}^{N}\theta_j X_{i,j},1\right) \text{ if } w_i=1 \qquad \text{(A.3)}$$

where the subscripts $-$ and $+$ indicate the negative $]-\infty,0[$ and positive $[0,\infty[$ regions of the support of the shifted standard normal distributions, respectively. In summary, for a binary phenotype, each round of the Gibbs sampling algorithm is started by sampling continuous latent variables y_i, after which the other Gibbs steps are the same as in the continuous phenotype case, with the exception that the variance σ_0^2 of the error terms is not updated, as it is now 1 by assumption.

Missing values: The analytical derivation of the fully conditional posterior distributions of the missing values is not feasible; therefore, we resort to the use of a Metropolis–Hastings step (Hastings, 1970; Chib and Greenberg, 1995), which includes first proposing a new candidate value, which is then either accepted or rejected. Thus, we proceed as follows:

1. Initialize all missing markers $m_{i,j}^{(0)}$ and gene expressions $x_{i,j}^{(0)}$ by sampling from their prior distributions.
2. At round t, one at a time, propose a new value for each missing marker or missing gene expression from the corresponding prior distribution. The proposed value is accepted with probability

$$\min\left(\frac{p(y_i|\theta,\mu,\sigma^2,D^{\text{prop}})}{p\left(y_i|\theta,\mu,\sigma^2,D^{(t-1)}\right)},1\right) \qquad \text{(A.4)}$$

otherwise it is rejected and the old value is retained. Here $D^{(t-1)}$ is the complete genetic data at round $t-1$ and $D^{(\text{Prop})}$ is the complete genetic data containing only one new proposed value (either genotype or gene expression level).

The calculation of the acceptance probability (A.4) can be made very efficient by considering its logarithm, which depends only on the few terms containing the missing value. For example, if marker $m_{i,j}$ is missing and it is not linked to any gene expression value, we only need to consider the difference $\alpha_j z_{i,j}^{\text{prop}} - \alpha_j z_{i,j}^{(t-1)}$ in the evaluation of the log likelihood ratio.