

ORIGINAL ARTICLE

Hierarchical modeling of clinical and expression quantitative trait loci

MJ Sillanpää and N Noykova

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

Previous articles have presented clinical quantitative trait locus (cQTL) models, where the information provided by quantitative/qualitative phenotypes, molecular markers and gene expressions (transcription levels) were combined and analyzed simultaneously. Because of financial constraints, marker data may be available for much larger group of individuals than expression data. However, it is desirable to use all the available information. We therefore extend such approaches by presenting a reliable missing data model for the case when marker data is more complete (that is, has many fewer missing entries). In the suggested hierarchical model, an expression QTL (eQTL) model (which is essentially our missing data model) is part of the larger cQTL model and it represents a Bayesian model-based method for estimating *cis*- and *trans*-acting regulatory effects for multiple (typically hundreds of) expression pheno-

types simultaneously. The modeling dependence between transcripts in the eQTL model is also considered. The method is based on presenting data in the form of marker gene pairs, for which the presence of regulatory effect (link) can be hypothesized. These marker gene pairs can be obtained from oligonucleotide arrays or created using information available on known pathways or previous eQTL/allelic expression studies. The estimation of the model parameters (such as presence/absence of regulation, eQTL/cQTL effects and proportion of eQTLs and cQTLs among the set of marker gene pairs) as well as the handling of missing data is performed using Markov Chain Monte Carlo (MCMC) sampling. The method is illustrated using both simulated and real data.

Heredity (2008) **101**, 271–284; doi:10.1038/hdy.2008.58; published online 23 July 2008

Keywords: eQTLs; cQTLs; Bayes; model selection; MCMC

Introduction

Expression quantitative trait locus (eQTL) studies (Jansen and Nap, 2001, 2004) have been conducted recently in man, mouse and other organisms (Schadt *et al.*, 2003; Morley *et al.*, 2004; Sladek and Hudson, 2006). In such studies both marker- and gene-expression data need to be available from each study individual. These studies utilize conventional QTL mapping to analyze genetic patterns (eQTLs) underlying the gene expressions and regulatory networks (Bystrykh *et al.*, 2005; Chessler *et al.*, 2005). A similar strategy has recently been applied also for studying genetic patterns of protein expression (Foss *et al.*, 2007). An eQTL could be *cis*- or *trans*-acting. *Cis*-acting means that the eQTL maps to the same (or a very close) genome position as the gene whose variation it explains. Similarly *trans*-acting eQTLs map to a distant genome locations to the genes and remotely regulate their expression. As in the case of expression profiling (Aune *et al.*, 2004), it is also possible to study colocalization of eQTLs with genome positions explaining clinical phenotype(s) (Mehrabian *et al.*, 2005; Schadt *et al.*, 2005).

The data collected for eQTL studies are two dimensional. The number of investigated gene transcripts

determines the first dimension (typically containing thousands of measurements), whereas the other dimension (typically containing dozens of measurements) refers to the number of genotyped individuals, forming the sample size. However, eQTL studies today still suffer from low reproducibility in microarray measurements (Draghici *et al.*, 2006) and small sample size in terms of individuals (de Koning and Haley, 2005), both of which give some cause for concern. To alleviate this problem, strategies for optimal design (Bueno Filho *et al.*, 2006) and selective phenotyping (Jin *et al.*, 2004; Jannink, 2005; Xu *et al.*, 2005; Fu and Jansen, 2006) have been developed.

Typically in an eQTL mapping study, data is screened over hundreds (or thousands) of different gene expressions (that is, expression phenotypes). The high dimensionality of the data may lead to serious computational problems. This encourages the use of some exploratory or preliminary screening methods or database information concerning interesting pathways, which are usually applied to reduce number of candidates (Thomas, 2005). The more detailed modeling efforts are then only targeted on the resulting subsets of data. The exception to this design are the marker-based approaches, where data at given marker point is first divided into genotypic subgroups and each subgroup is then searched for differentially expressed genes by using standard methods (see Kendziorski *et al.*, 2006; for a review see Parmigiani *et al.*, 2003). However, all of these approaches suffer from certain flaws. The repeated application of statistical test leads to serious concerns about the

Correspondence: Dr MJ Sillanpää, Department of Mathematics and Statistics, Rolf Nevanlinna Institute, P.O. Box 68, FIN-00014, University of Helsinki, Helsinki, Finland.

E-mail: mjs@rolf.helsinki.fi

Received 13 December 2007; revised 13 May 2008; accepted 23 May 2008; published online 23 July 2008

appropriate significance threshold due to dependency between the tests and the multiple testing problem. The subset of differentially expressed genes may not necessarily represent a functionally important subset of genes (Yanai *et al.*, 2006). Moreover, the selection of candidates based on dimension reduction techniques (Perez-Enciso *et al.*, 2003; Lan *et al.*, 2004) has difficulties concerning the interpretation of the new variables. And finally, interesting pathways may still include hundreds of genes that necessitate the development of new effective analysis methods. For some new developments, see Chen and Kendzierski (2007), Gelfond *et al.* (2007), Jia and Xu (2007), Perez-Enciso *et al.* (2007).

The same kind of data (markers and expression jointly) can be used to explain variation in quantitative traits, which is called clinical quantitative trait locus (cQTL) analysis (Hoti and Sillanpää, 2006; Bhattacharjee and Sillanpää, 2008). It is evident in this setup that the expression measurements of the joint data can provide additional information for explaining and predicting the phenotype (West *et al.*, 2006). Due to financial constraints the current problem in cQTL analysis is having a too small sample size for the joint data even if marker data may be available for much larger group of individuals. Here we want to address the problems of predicting expression values for genotyped individuals by integrating the eQTL model, as a missing data model, into the cQTL model. Thus, we present method to estimate parameters underlying the frequency distribution of gene expression among the prespecified set of marker gene pairs. Information from previous eQTL and allelic expression studies as well as known pathways can be utilized in the forming of such input data (marker gene pairs; Hoti and Sillanpää, 2006). The suggested method provides posterior estimates and predictions for parameters (for example, missing data) including: (1) the proportions and occupancy probabilities of the eQTLs (the markers regulating the expression) and the cQTLs (the marker or transcript variation explaining the phenotype) as well as their eQTL and cQTL effects, (2) the predicted values of gene expression based on the genotypes at a regulatory locus and (3) the genotype predictions based on the expression values and the genotypes at linked (adjacent) loci. Because of the above listed properties the suggested method can also be regarded as a multi-trait eQTL analysis, which can simultaneously handle hundreds of expression phenotypes.

The model

We first introduce the eQTL model for molecular marker and expression data and then present a large hierarchical model for quantitative phenotypes, the cQTL model, as an extension of our eQTL model.

Input data

Let us assume that offspring data from inbred line cross experiment (backcross, double haploids or F_2) consist of paired marker- and gene-expression measurements, marker-gene pairs, collected from each individual separately. Such data was called link data in Hoti and Sillanpää (2006) and is here considered to represent earlier eQTL findings from allele expression studies, genetical genomics experiments or known pathways that are to be validated. Both *cis*- and *trans*-acting pairs can be

included but the validation data set cannot be the same set where the original findings were made (to avoid selection bias). However, the *trans*-acting effects are known to be small and thus more difficult to identify (Ren *et al.*, 2000; Sladek and Hudson, 2006). Moreover, *trans*-acting eQTLs often occur in clusters (Mueller *et al.*, 2006). Therefore we focus mainly on the larger *cis*-acting effects, which could be successfully identified from the current small data sets in presence of missing data. If there are no earlier findings, the suitable data for the method would be oligonucleotide array data of Ronald *et al.* (2005) where the marker- and the gene-expression measurements are simultaneously produced at every position. Alternatively, to study *cis*-acting regulation, one can form putative link data (presented as pairs) solely based on the genomic proximity between the markers and the genes.

Here we assume that only a single marker (a major gene/major-effect eQTL) is controlling each expression phenotype, which means that a gene-expression distribution can depart from a normal distribution (Gibson and Weir, 2005). Note, however, that under this assumption, the same marker can simultaneously regulate two or more expressions. See 'Discussion' for multilocus modeling of expression phenotype.

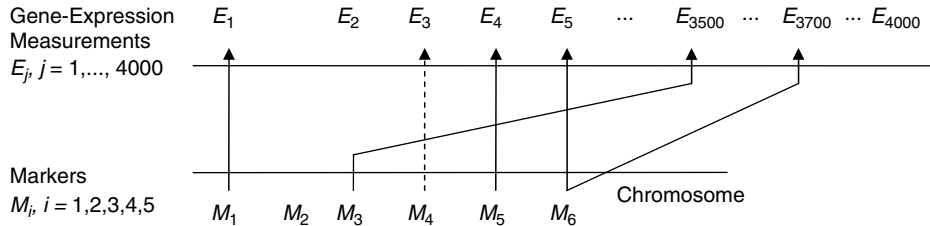
The *cis*- and *trans*-acting effects and the corresponding marker gene pairs are illustrated in Figure 1. On the same figure the form and indexing of the input data, which are chosen *cis*- and *trans*-acting pairs, is described. This indexing is required as a first step of the statistical data description and follows the order of the markers on chromosomes. In case there is no information about the expression value, related to some particular marker, this marker is included in the input data as a pair with missing information about the corresponding gene expression. In case when two gene expressions are both regulated by the same marker we assume that this marker is represented twice, so that the distance between the both copies is specified to be extremely small (approximately 0).

Expression QTL model

Let us assume that backcross or double haploid data has been collected from N individuals at N_p marker gene pairs. See Appendix for consideration of F_2 intercross. For a convenience, two genotypes are denoted as AA and Aa in case of backcross; AA and aa for double haploid data. Conditionally on the underlying parameters explained below the following bimodal mixture distribution is assumed for the expression data, where i is index for an individual ($i = 1 \dots N$), and j for a marker gene pair ($j = 1 \dots N_p$):

$$E_{i,j} | I_j, \mu_j, A_j, G_{i,j}, \alpha_j, \sigma_j^2 \sim N(\alpha_j + I_j \mu_j A_j G_{i,j}, \sigma_j^2) \quad (1)$$

This is equivalent to assuming (simultaneously for each pair) a linear eQTL model $E_{i,j} = \alpha_j + I_j \mu_j A_j G_{i,j} + \varepsilon_{i,j}$, where the residuals $\varepsilon_{i,j}$ (that is, the expression values after correcting with respect to the regulatory effects) follow a normal distribution with mean 0 and variance σ_j^2 . After normalization and transformation of the data (Quackenbush, 2001) we assume that the overall mean and the expression variance in each pair and in each mixture component are equal: $\alpha_j = \alpha_0$ and $\sigma_j^2 = \sigma_0^2$ for all j . Moreover, we assume that data has been centralized,



Listing of the pairs

Cis- pairs: $\{(M_1, E_1), (M_4, E_3), (M_5, E_4)\}$

Trans- pairs: $\{(M_3, E_{3500})\}$

Pairs with missing expression data: $\{(M_2, \bullet)\}$

Pairs, where single marker regulates several expressions: $\{(M_6, E_5), (M_6, E_{3700})\}$

Correspondence of the ordering of the pairs in the input data to the ordering of the pairs in the original data:

$\{(M_1, E_1), (M_2, E_2), (M_3, E_3), (M_4, E_4), (M_5, E_5), (M_6, E_6), (M_7, E_7)\}$
 $= \{(M_1, E_1), (M_2, \bullet), (M_3, E_{3500}), (M_4, E_3), (M_5, E_4), (M_6, E_5), (M_6, E_{3700})\}$

Figure 1 An example of ordering of known *cis*- and *trans*-acting marker gene pairs. In this example 4000 expression measurements and information from six markers are available. On the basis of the previous independent experiment only three *cis*-acting pairs with clear one-to-one correspondence and two *trans*-acting effect pairs are expected. There is no prior information about *cis*-acting effects between the marker M_4 and the gene-expression E_3 , but this putative pair (M_4, E_3) is included in the statistical analysis because M_4 and E_3 are very close to each other on the genome. There is no information about any E_j expression, connected with the marker M_2 . Therefore (M_2, \bullet) is a pair with missing information. There are two expressions (E_5 and E_{3700}), both corresponding to the marker M_6 . In this case the pairs $\{(M_6, E_5), (M_6, E_{3700})\}$ are formed. The ordering of all pairs in the input data follows the chromosomal ordering of the markers.

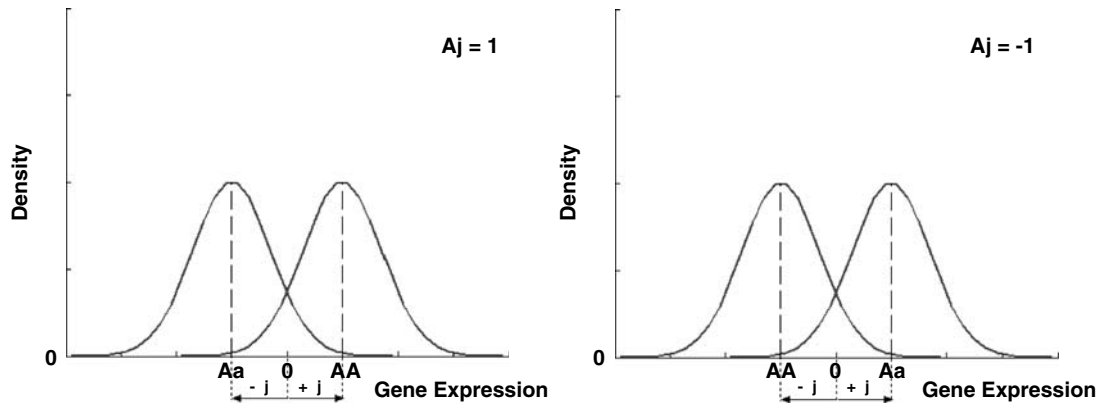


Figure 2 Backcross. The gene-expression distribution for the possible values of the assignment variable A_j , corresponding to the left ($A_j = 1$) and the right ($A_j = -1$) ordering of the genotypes Aa and AA . Here $A_j = 1$ means that there is a positive regulatory effect, and $A_j = -1$ means a negative regulatory effect.

that is, $\alpha_j = \alpha_0 = 0$, and the residuals $\varepsilon_{i,j}$ are uncorrelated even if centralization may induce dependence between residuals in practice (Qu and Xu, 2006; Jia and Xu, 2007). Here, the value of the indicator variable I_j controls presence ($I_j = 1$) or absence ($I_j = 0$) of a regulatory effect for pair j . A variable $\mu_j \geq 0$ is the effect size, and a variable $G_{i,j}$ is the genotype value of individual i at marker j , which is 1 for genotype AA and -1 for the other genotype. The value of the assignment variable A_j for pair j defines the sign of the regulatory effect: $A_j = 1$ corresponds to the positive and $A_j = -1$ to the negative effect (Figure 2).

Note that, in case of backcross, one can alternatively learn the values of the assignment variable A_j by taking one extra microarray from pooled sample of individuals from one of the parental lines (all with genotype AA) and possibly one from F_1 individuals (Aa). From Figure 2, it becomes clear that an individual's genotype (at regulatory locus) can be predicted by knowing only the values of the assignment and gene expression. This immediately suggests one strategy to produce genotype predictions for known marker gene pairs based on the gene expressions from the offspring and one of the parents.

Hence the expression data $E_{i,j}$ are described as a mixture of two normal distributions, centered around $-\mu_j$ and $+\mu_j$ and depending on the possible values of genotypes AA and Aa . Thus the gene effect (assuming co-dominance) is presented by the quantity $2I_j\mu_j$, where the product $\beta_j = I_j\mu_j$ is called as a regulatory effect of the gene.

In the case where we allow the effect size μ_j to be also negative, $\mu_j \in (-\infty, +\infty)$, and fix the assignment variable $A_j = 1$, we practically obtain the same model (1), described above. Then the information from the assignment variable A_j is in the sign of μ_j because only the product $\mu_j^* = \mu_j A_j$ is involved in the description of the bimodal distribution (1). The separate sign variable A_j is simply used here for illustration and to emphasize that knowledge of it may potentially be useful for genotype prediction.

Hierarchical eQTL model

Denote the data vector as $D = (E^O, G^O)$, where the observed gene expression (E^O) and marker data (G^O) both may have some missing entries. The eQTL-parameter vector is denoted as $\theta^e = (I, \mu, A, G, E, \sigma_0^2)$, where E and G represent the complete forms of the data. The mutual independence is assumed between and among the variables I , μ and A . According to Bayes rule, $p(\theta^e | D) = p(D, \theta^e) / p(D) = c \cdot p(D, \theta^e)$, where $c = 1/p(D)$ is a normalizing constant. Here the posterior distribution $p(\theta^e | D)$ is proportional to the joint distribution of the parameters and data, $p(D, \theta^e)$, which can be expressed as a product of likelihood $p(D | \theta^e)$ and prior $p(\theta^e)$. This is equivalently $p(\theta^e | D) \propto p(D, \theta^e) = p(D | \theta^e) p(\theta^e)$ and, for given conditional independence assumptions, can be further factorized as:

$$p(I, \mu, A, G, E, \sigma_0^2 | E^O, G^O) \propto p(E^O, G^O, I, \mu, A, G, E, \sigma_0^2) \\ = p(E^O, G^O | E, G) p(E | I, \mu, A, G, \sigma_0^2) p(A) p(G) p(\mu) p(I | s^e) p(\sigma_0^2).$$

Here $p(E^O, G^O | E, G)$ is the indicator function being one only when the complete data is consistent with the observations, and is 0 otherwise. The complete expression data likelihood

$$p(E | I, \mu, A, G, \sigma_0^2) = \prod_{i=1}^N \prod_{j=1}^{N_p} p(E_{i,j} | I_j, \mu_j, A_j, G_{i,j}, \sigma_0^2),$$

where the likelihood function can be written for individual i and pair j (equation 1) as:

$$p(E_{i,j} | I_j, \mu_j, A_j, G_{i,j}, \sigma_0^2) \\ = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{1}{2\sigma_0^2} (E_{i,j} - I_j\mu_j A_j G_{i,j})^2\right).$$

Prior distributions: Next we define functional forms of prior distributions, reflecting our prior beliefs, for parameters $p(A)$, $p(G)$, $p(\mu)$, $p(I | s^e)$, and $p(\sigma_0^2)$. We assume that the assignment variables A_j are mutually independent and have the following prior distribution, $p(A) = \prod_{j=1}^{N_p} p(A_j)$, where $p(A_j)$ is a Bernoulli (π_j) distribution with parameter $\pi_j = \frac{1}{2}$ at each locus. This means that both assignments for A_j are *a priori* equally likely. The prior density function of the effect size can be expressed as $p(\mu) = \prod_{j=1}^{N_p} p(\mu_j)$, where $p(\mu_j)$ is a density

function of right (positive) tail of normal distribution (truncated at 0) with mean 0 and variance 100. The prior for indicator variables I is $p(I | s^e) = \prod_{j=1}^{N_p} p(I_j | s^e)$, where $p(I_j | s^e)$ is a Bernoulli (s^e) distribution with known parameter $s^e = P(I_j = 1)$. The parameter $s^e \geq \frac{1}{2}$ represents our prior expectation for the proportion of pairs with regulatory effect, that is, it controls how much value one is preferred over 0. Its value is assumed to be greater than $\frac{1}{2}$ because it is probable that a large proportion of pairs (to be validated) actually has a regulatory effect. The prior $p(\sigma_0^2)$ is assumed to be an inverse Gamma (1,1) restricted to the range (0.5, 10000) (cf. Sillanpää and Bhattacharjee, 2005, 2006). For discussion of alternative priors, see Gelman (2006) and Van Dongen (2006). Note that the restriction of the inverse Gamma distribution was imposed for computational reasons—to maintain numerical stability in OpenBUGS.

Model for missing genotypes: The prior distribution of the marker data is defined in the same way as in Sillanpää and Arjas (1998) and Hoti and Sillanpää (2006). We assume that the genotype measurements are conditionally independent between individuals (given the parents), because all individuals are equally related:

$$p(G) = \prod_{i=1}^N p(G_{i,1}, G_{i,2}, \dots, G_{i,N_p}) \\ \propto \prod_{i=1}^N \left[p(G_{i,1}) \prod_{j=1}^{N_p} p(G_{i,j} | G_{i,j-1}) \right],$$

where $P(G_{j,1})$ is the prior probability (expected frequency) of genotype $G_{i,1}$ at marker 1 and $p(G_{i,j} | G_{i,j-1})$ is the between-loci transition probability for individual i . The actual values depend on the genotypes, the map distance (the recombination fraction) and the design considered (for details, see Jiang and Zeng, 1997; Sillanpää and Arjas, 1998). In case where the genetic map is unknown, unlinked loci and expected genotype frequencies can be assumed in $p(G) = \prod_{i=1}^N \prod_{j=1}^{N_p} p(G_{i,j})$.

Model selection and interpretation: Bayesian model selection of pairs with a regulatory effect is performed using indicator variables. A similar technique is commonly used for variable selection in QTL and association models (Uimari *et al.*, 1996; Uimari and Hoeschele, 1997; Yi *et al.*, 2003; Sillanpää and Bhattacharjee, 2005, 2006). According to the above prior assumptions, the parameters μ_j and I_j are *a priori* independent, that is, $p(\mu, I | s^e) = p(\mu) p(I | s^e)$. This formulation is analogous to Kuo and Mallick (1998). Thus, based on evidence given by data, to conclude if pair j has a regulatory effect, it is more robust to monitor the posterior product $\beta_j = \mu_j \times I_j$ rather than each variable (μ_j or I_j) separately (Sillanpää and Bhattacharjee, 2005). Alternatively one can assume a hierarchical prior $p(\mu, I | s^e) = p(\mu | D) p(I | s^e)$, which gives better identifiability for individual parameters but additional computational problems such as adjustment of tuning parameters (pseudo priors) may appear unless μ_j and I_j are updated together as a block during the Markov Chain Monte Carlo (MCMC) sampling (Geweke, 1996; Meuwissen *et al.*, 2001).

Modeling dependence between transcripts: So far we have assumed independence between expression levels at different pairs. However, the values of gene expressions may be (in reality) correlated due to many different reasons. In the following, we hypothesize two kinds of dependencies: (1) the spatial dependencies due to genomic proximity of the gene transcripts meaning that the expression distributions of two genes, whose positions are close by in the genome, are dependent on each other according to the distance between them and (2) the dependencies due to the membership of the genes in the same pathway/gene set. To consider dependence between expressions we model it in the level of their underlying distributions, at effect sizes μ_j . That is, we model the effect size vector $\boldsymbol{\mu} = (\mu_1, \dots, \mu_{N_p})^T$ using a multivariate normal distribution, $\boldsymbol{\mu} \sim \text{MVN}(\boldsymbol{O}, \mathbf{S})$ with the mean vector $\boldsymbol{O} = (0, 0, \dots, 0)^T$ and the covariance matrix \mathbf{S} .

The existence of spatial dependence between expression distributions may not be well justified biologically, but it provides helpful way to share information horizontally between transcripts. In the spatial dependence model, the elements of the covariance matrix are given by the exponential decay function $s_{j,j+1} = \tau_0 \exp(-\lambda d_j)$, which depends on the given smoothing parameters (τ_0 and λ) and the physical or genetic distance d_j between the transcripts j and $j+1$. The parameter τ_0 controls the overall level of smoothing (say $\tau_0 = 100$) and λ defines the degree of spatial dependence (Conti and Witte, 2003; Sillanpää and Bhattacharjee, 2005). Note that this model is especially suitable for densely spaced transcripts (spanning few cM candidate regions) because dependence is a decreasing function of genomic distance and the rate is dependent on τ_0 and λ (cf., Sillanpää and Bhattacharjee, 2005).

In the pathway membership model, the connectivity matrix \mathbf{S} (with all elements being τ_0 or 0) is constructed based on the database knowledge about the pathway memberships, list of differentially expressed genes or simply based on the pairwise correlateness between the gene expressions. In the last case, if linear pairwise expression correlation between the genes j and k is higher or equivalent than a predefined threshold T , then two genes are said to be connected, that is, if $\rho_{j,k} \geq T$ then $s_{j,k} = \tau_0$ and otherwise $s_{j,k} = 0$. The parameter τ_0 is the prior variance/covariance assumed for the effect size among the pathway members.

Clinical QTL model

Hoti and Sillanpää (2006) presented a cQTL model where a phenotype $Y = (Y_i)$ was described as a linear combination of the marker genotypes $G = (G_{i,j})$ and the gene-expression levels $E = (E_{i,j})$ and possible genotype \times expression interactions. The genotype \times expression interactions are allowed to occur only between members (genotypes and expressions) of the single marker gene pair. Due to necessary assumption of co-dominance in backcross, these genotype \times expression interactions should be interpreted as allele-specific expression effects. Here we use the similar model than Hoti and Sillanpää (2006) except that the phenotype-associated subset of terms is determined by the indicator variables (cf. Bhattacharjee and Sillanpää, 2008; see ‘Model selection and interpretation’ above). The generic term cQTL is used for the trait-associated components. We assume the

following cQTL model for the quantitative phenotype Y_i of individual i :

$$Y_i = a + \sum_{j=1}^{N_p} \left(I_j^M \beta_j^M G_{i,j} + I_j^E \beta_j^E E_{i,j} + I_j^{ME} \beta_j^{ME} G_{i,j} E_{i,j} \right) + e_i. \quad (2)$$

Here a is an overall mean and the residuals e_i (=observed–estimated trait value) are assumed to be normally distributed with mean 0 and variance σ_e^2 . Let us denote an indicator variable for the marker and the transcript at each pair j as I_j^M and I_j^E , respectively. Similarly, let us denote an indicator variable for the genotype-expression interaction component (at pair j) as I_j^{ME} . These indicator variables are together collected into the single vector of triplets as $I^c = (I_1^M, I_1^E, I_1^{ME}, \dots, I_{N_p}^M, I_{N_p}^E, I_{N_p}^{ME})^T$. For pair j , the genotype, expression, and genotype \times expression interaction effects with respect to phenotype are determined by β_j^M , β_j^E and β_j^{ME} . The cQTL effects are jointly denoted in the vector form as $\beta^c = (\beta_1^M, \beta_1^E, \beta_1^{ME}, \dots, \beta_{N_p}^M, \beta_{N_p}^E, \beta_{N_p}^{ME})^T$. For F2 intercross, see Appendix. To model binary phenotypes, see Hoti and Sillanpää (2006) and Bhattacharjee and Sillanpää (2008).

Hierarchical cQTL model

Let us denote the cQTL model parameters as $\theta^c = (I^c, \beta^c, \sigma_e^2, a, \sigma_a^2)$ and recall that the eQTL model parameters were denoted as $\theta^e = (I, \mu, A, G, E, \sigma_\mu^2)$. Now their posterior distribution is proportional to the joint distribution (of data and parameters) and can be further factorized as

$$\begin{aligned} p(\theta^c, \theta^e | E^O, G^O, Y) &\propto p(\theta^c, \theta^e, E^O, G^O, Y) \\ &= p(Y | a, I^c, \beta^c, E, G, \sigma_e^2) p(I^c | s^c) p(\beta^c | \sigma_e^2) \\ &\quad \times p(\sigma_e^2) p(a) p(\sigma_a^2) \times p(E^O, G^O, \theta^e) \end{aligned}$$

Here the likelihood function for all individuals jointly is

$$\begin{aligned} p(Y | a, I^c, \beta^c, E, G, \sigma_e^2) &= \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left(-\frac{1}{2\sigma_e^2} \left(Y_i - a \right. \right. \\ &\quad \left. \left. - \sum_{j=1}^{N_p} \left(I_j^M \beta_j^M G_{i,j} + I_j^E \beta_j^E E_{i,j} + I_j^{ME} \beta_j^{ME} G_{i,j} E_{i,j} \right) \right)^2 \right). \end{aligned}$$

We make several (conditional) independence assumptions in the construction of prior distributions. Given $s^c = p(I^c = 1)$, which is a small prior probability for a candidate to be associated into the trait, we assume the following independence prior for the indicator variables $p(I^c | s^c) = \prod_{j=1}^{N_p} [p(I_j^M | s^c) \times p(I_j^E | s^c) \times p(I_j^{ME} | s^c)]$. Here $p(I_j^k | s^c)$ for each component j and $K = \{M, E, ME\}$ is a Bernoulli (s^c) distribution with parameter s^c . Note that unlike the s^c of eQTL model, we consider $s^c \leq \frac{1}{2}$ to be very small. Similarly, we assume a prior for the genetic effects $p(\beta^c | \sigma_e^2) = \prod_{j=1}^{N_p} [p(\beta_j^M | \sigma_{M(j)}^2) \times p(\beta_j^E | \sigma_{E(j)}^2) \times p(\beta_j^{ME} | \sigma_{ME(j)}^2)]$, where $p(\beta_j^K | \sigma_{K(j)}^2)$ for each (coefficient at) component j and $K = \{M, E, ME\}$ is a normal distribution with mean 0 and variance $\sigma_{K(j)}^2$. The prior for genetic variances is assumed to be $p(\sigma_e^2) = \prod_{j=1}^{N_p} [p(\sigma_{M(j)}^2) \times p(\sigma_{E(j)}^2) \times p(\sigma_{ME(j)}^2)]$ and $p(\sigma_{K(j)}^2)$ for each component j and $K = \{M, E, ME\}$ is an inverse Gamma (1,1) without any boundaries. The prior $p(a)$ is assumed to be flat normal distribution with mean 0 and variance 10 000. The prior for residual variance $p(\sigma_e^2)$ is assumed to be inverse Gamma (1,1) restricted to

the range (0.5, 10 000). It is useful to note that our earlier eQTL model constitutes a missing data model within this large hierarchical cQTL model specified above. Moreover, the likelihood of the eQTL model is the prior in the cQTL model.

Applications

Before presenting analyses using the complete cQTL model with simulated data, we focus clearly on eQTL model as independent model structure. Thus, in following, we first present several trials by using only eQTL model in our analyses. With simulated (complete) data, we consider both the performance under a single realization of a data set and the average performance by analyzing 50 data replicates. In addition, we consider the performance of two realizations of data sets in presence of missing values. Then we present the eQTL model analyses using previously analyzed real double haploid data on *Saccharomyces cerevisiae* (Brem *et al.*, 2002) in the original and transformed scales as well as the accuracy assessment of the predicted expression values. Finally, we study performance of eQTL model with simulated data in presence of dependence between transcripts. For data simulation and for the MCMC estimation of eQTL model (1) parameters in these experiments, we have systematically used the OpenBUGS 2.2.0 software (Spiegelhalter *et al.*, 2005; Thomas *et al.*, 2006) if not stated otherwise. In the analyses, the first 10 000 MCMC iterations were discarded from the chain as 'burn-in'. The posterior estimates are based on the next 100 000 MCMC iterations. To summarize the results of continuous parameters, we have preferred to use posterior median instead of posterior mean (available in OpenBUGS) as our point estimate approximating posterior mode (Hazelton and Gurrin, 2003). For discrete parameters (and product of discrete and continuous parameters), we have used the posterior mean. In eQTL analyses, the key criterion of assessment of performance has been the estimation and the prediction error.

Simulated eQTL data

Simulating markers: The linked marker data G (over 102 marker points) was simulated using the WinQTL Cartographer program (Wang *et al.*, 2006) for 200 backcross individuals resulting from an inbred line cross experiment. The marker data spanned three chromosomes of length 99 cM, so that there were 34 evenly spaced markers on every chromosome. The distance between every two markers was 3 cM.

Simulating expressions: The gene-expression value $E_{i,j}$ for each individual i and for each marker gene pair j , was simulated in the OpenBUGS conditionally on the marker data $G_{i,j}$ and the parameters according to the eQTL model (1). At each locus j , the selection indicator I_j was generated from a Bernoulli distribution, with Bernoulli parameter $s^e = P(I_j = 1) = 0.9$, which means that majority (90%) of the pairs are likely to have regulatory effect. The residual variance was set to $\sigma_0^2 = 1$ and the overall mean to $\alpha_0 = 0$. For every pair j , the effect size μ_j was generated from a truncated flat normal distribution with mean 2 and variance 100, which was restricted to the range [0,4]. Thus only moderate positive values of μ_j were possible.

All regulatory effects were set to be positive by having $A_j = 1$ for each pair j .

Most of the pairs (G, E) in the simulated data follow the typical bimodal gene-expression frequency distribution (Figure 3a). In some cases (Figure 3b) the resulting frequency distribution does not strictly follow the bimodal shape because of some overlapping between the two mixture components.

In the cases when μ_j is near to 0, the tails of the two mixture components of E_j almost completely overlap (Figure 3c).

Analysis of the single simulated eQTL data set: Note that the known values of $\sigma_0^2 = 1$ and $\alpha_0 = 0$ were assumed in the analysis of the simulated backcross data. The weakly informative prior (a truncated normal distribution) was considered for the effect size $\mu_j \sim N(2, 100)$ with the restriction $[0, \infty[$. From the eQTL model (1) it becomes clear that when there is no regulatory effect ($\beta_j = 0$), the parameter A_j does not have any interpretation. Excluding such positions, the estimated values of A_j match perfectly well with the true simulated values. In Figure 4a, based on the posterior estimates (the median β_j^{med} and 95% credible interval) for the regulatory effect $\beta_j = I_j \times \mu_j$ of pair j , we present the estimation error $q_j = \beta_j^{med} - \beta_j^i$ (such as, a deviation from the true simulated value β_j^i) and the corresponding credible interval as a summary of the analysis.

Analysis of 50 simulated eQTL data replicates: Next, 50 replicated data sets (simulated replicates) of size $N = 200$ were simulated using the same generating eQTL model as above. Every data set ($d = 1, \dots, 50$) was analyzed using OpenBUGS similarly as above. For data set d , let us denote the posterior median of the estimated regulatory effect at position j as $\beta_j^{med}(d)$ and its estimation error as $q_j(d) = \beta_j^{med}(d) - \beta_j^i(d)$ (viz. a deviation from the true simulated value $\beta_j^i(d)$). Figure 4b presents the mean $M_j(q_j)$, the median $m_j(q_j)$, and the standard deviation $SD_j(q_j)$ of the estimation error for every marker j over 50 simulation replicates (note the scale of the y -axis). These summaries indicate that the estimation errors are very small supporting the conclusion that our method can provide reliable eQTL effect estimates (when σ_0^2 and α_0 are known).

Analysis of the single simulated eQTL data set in presence of missing data: To check the sensitivity of the method to randomly occurring missing values we have analyzed here the same simulated backcross data (explained above) in two different cases: (1) when 10% of backcross data were coded (in random locations) as missing among both the marker ($G_{i,j}$) and the expression ($E_{i,j}$) measurements, and (2) when 10% of the marker data ($G_{i,j}$) and 50% of the expression data ($E_{i,j}$) in the simulated backcross were coded (in random locations) as missing.

The two data sets with missing values were analyzed using OpenBUGS similarly as above. We assume that values of the outcome variable (here the gene expressions) are missing at random. This is a default assumption in OpenBUGS implying that the posterior distributions of the eQTL model parameters are influenced only by the observed part of the outcomes E^o (Rubin, 1976).

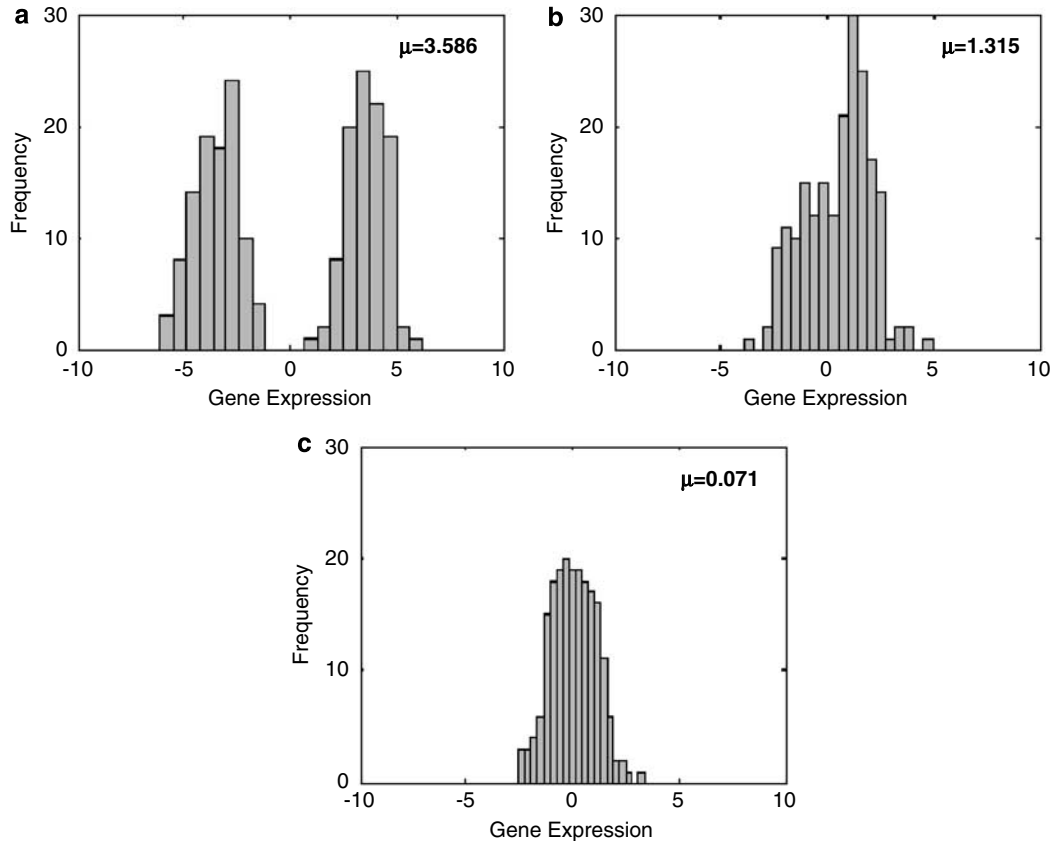


Figure 3 Three typical cases of the simulated gene-expression (E_j) data: (a) the well-distinguished bimodal frequency distribution when μ_j is high ($\mu_j = 3.586$); (b) the case when the simulated E_j data components overlap, but it is still possible to distinguish two parts ($\mu_j = 1.315$); (c) two parts of the distribution overlap almost completely when μ_j is small ($\mu_j = 0.071$). The expression values are shown on the x-axes and the frequencies on the y-axes.

From Figure 4c, it becomes clear that the obtained estimates are reliable in presence of 10% missingness. The same conclusion is valid also in the case when 50% of the expression measurements and 10% of the marker data are missing (Figure 4d). As expected, the credible interval (for the estimation error) becomes wider with increasing amount of missingness. Although it is visible in Figure 4d that the posterior median has a larger amplitude than in A and C, we found out that this was not the case for the relative estimation error (calculated for non-zero coefficients), which practically stays a constant level at these three cases (results not shown). Thus the results suggest that the performance of the method is generally robust to the presence of missing observations.

Real yeast data

We selected the publicly available data from a double haploid experiment on *S. cerevisiae*, described in Brem *et al.* (2002), and used as test data in Hoti and Sillanpää (2006). The data contain the gene expressions (a dye swap pair of arrays) and the marker genotypes measured from 40 individuals (segregant samples) obtained from a cross between a laboratory (BY4716) and a wild strain of Yeast. The expression data represent the background corrected and normalized log ratios, which have been centered over all the microarray spots, that is, the mean of the expression data is 0. For each gene we took simply an average of two expression values (a dye swap pair of gene expressions) if both values were available and

marked it as missing otherwise; this procedure did not significantly change the centering from 0.

Analysis of yeast data using eQTL model: Brem *et al.* (2002) found 570 eQTLs, which we simply took together with the appropriate expressions as our input data (marker gene pairs) here. (Note that we are aware of the potential selection bias that may appear as a consequence of using data twice.) To validate these eQTLs, the four different analyses with the eQTL model were executed for the data: (1) the eQTL analysis of the original data assuming the known values of $\sigma_0^2 = 1$ and $\alpha_0 = 0$, (2) the eQTL analysis of the transformed data (the expression values of each gene were rescaled to have an unit variance by the common scaling factor) again assuming the known values of $\sigma_0^2 = 1$ and $\alpha_0 = 0$, (3) the eQTL analysis as 1 above with unknown σ_0^2 and (4) the eQTL analysis of the transformed data (the expression values of each gene were rescaled to have an unit variance by the locus-specific scaling factor) assuming the known values of $\sigma_0^2 = 1$ and $\alpha_0 = 0$. The uninformative prior (a uniform distribution; unlinked loci) was assumed for the genotype data in $p(G)$. In addition, two different priors (truncated normal distributions) were considered for the effect size in all four analyses: $\mu_j \sim N_+(0, 100)$ (a neutral prior) and $\mu_j \sim N_+(2, 100)$ (a nearly neutral prior), where the subscript $+$ indicates the positive $[0, \infty[$ region of support. This resulted in the eight different analyses in total. A data transformation was performed using

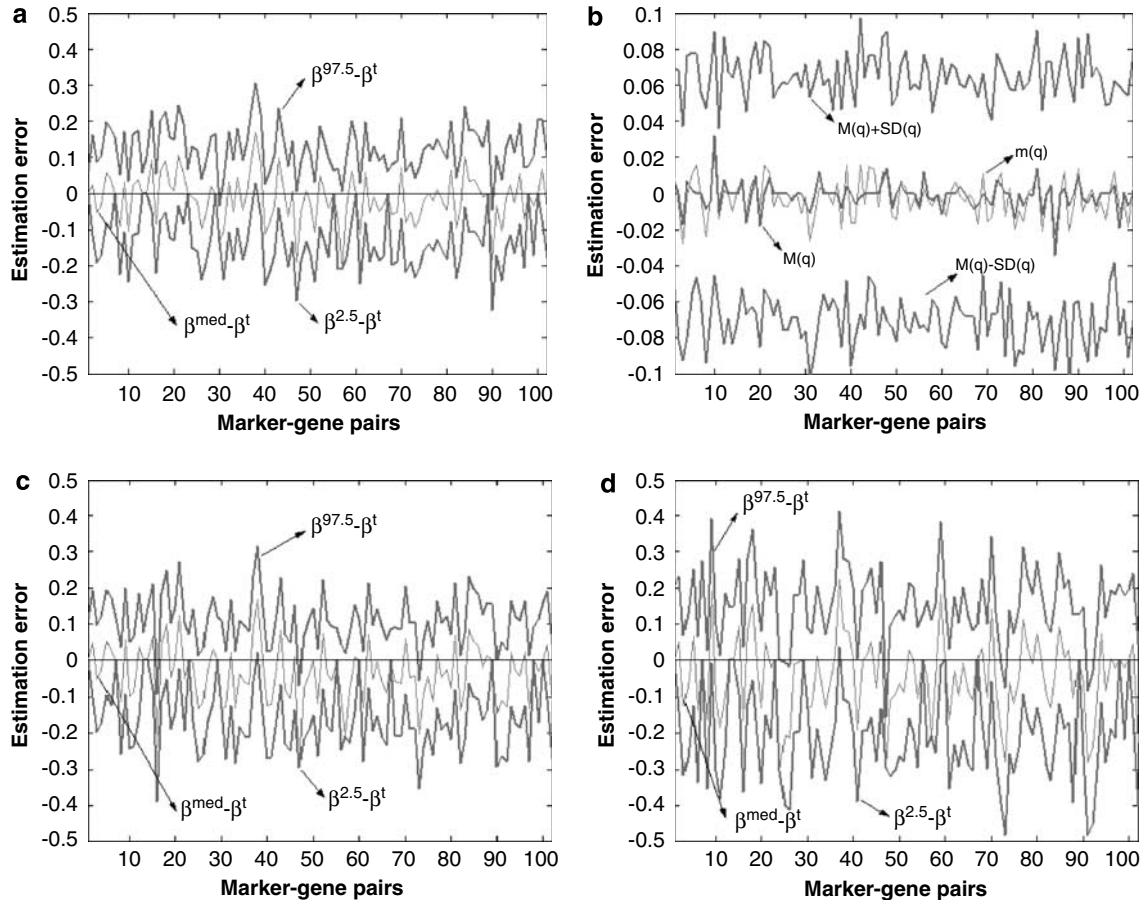


Figure 4 Summary of the estimation error of the posterior estimated regulatory effect $\beta_j = I_j \times \mu_j$ over 102 marker gene pairs. (a) The analysis of the single simulated data set. (b) The analysis of 50 simulation replicates. (c) The analysis of the single simulated data set with 10% of both the markers and the expressions missing. (d) The analysis of the single simulated data set with 10% of the markers and 50% of the expressions missing. In (a, c, and d) the estimation error (of the posterior median and 2.5 and 97.5% quantiles) around the true simulated value β_j is presented at each position j . In (b) the mean $M_j(q_j)$, the median $m_j(q_j)$ and the standard deviation (around the mean) $M_j(q_j) \pm SD_j(q_j)$ of the estimation error are shown for every marker gene pair j over data sets. The pairs are shown on the x-axis and the estimation error on the y-axis.

the formula $E_j^{st} = E_j^{data} / \hat{\sigma}_0$ in analysis 2 and the formula $E_j^{st} = E_j^{data} / \hat{\sigma}_{E_j}$ in analysis 4. Here $\hat{\sigma}_0 \approx 1.29$ is the common empirical standard deviation of all the expressions and $\hat{\sigma}_{E_j}$ is the empirical standard deviation of the expressions at gene j . By including the missing outcomes in the OpenBUGS analysis, one obtains the posterior predicted values for them based on the posterior distributions of the parameters. In Table 1, we show the posterior (median) estimated proportion of eQTLs, $P(\sum I_j / 570 | data)$, for all eight analyses corresponding to different prior assumptions for $s^e = P(I_j = 1)$. In the table, the observed proportion of non-zero posterior (median) estimated regulatory effects is also shown. The Monte Carlo error of the quantity $\sum I_j$ was estimated to be around 0.12 and 0.13 resulting to very accurate estimates for $\sum I_j / 570$, that is, error around $0.12 / 570$. The Monte Carlo error for β_j varied and was usually smaller than 0.003 but to estimate the observed proportion, this error should be multiplied with the number of non-zero positions. In other words, the posterior proportions, $P(\sum I_j / 570 | data)$, are much more accurate than the observed proportions in Table 1.

In Table 1, the analyses with the original data and the transformed data (by a common scaling factor) seem to

Table 1 Summary of the posterior and the observed proportions of eQTLs for different values of prior proportion $s^e = P(I_j = 1)$ in four analyses of yeast data (which each are evaluated at two priors of μ): the original data with the known $\sigma_0^2 = 1$ (top left), the transformed data (by using a common scaling factor) with the known $\sigma_0^2 = 1$ (top right), the original data with unknown σ_0^2 (bottom left), the transformed data (by using a gene-specific scaling factor) with the known $\sigma_0^2 = 1$ (bottom right)

Prior s^e	Posterior proportion				Observed proportion			
	$\mu \sim N_+(0,100)$	$\mu \sim N_+(2,100)$	$\mu \sim N_+(0,100)$	$\mu \sim N_+(2,100)$	$\mu \sim N_+(0,100)$	$\mu \sim N_+(2,100)$	$\mu \sim N_+(0,100)$	$\mu \sim N_+(2,100)$
0.75	0.444	0.388	0.428	0.374	0.374	0.319	0.363	0.298
	0.486	0.851	0.470	0.840	0.440	0.870	0.421	0.856
0.80	0.447	0.393	0.458	0.402	0.405	0.342	0.386	0.325
	0.517	0.869	0.500	0.858	0.456	0.882	0.451	0.875
0.85	0.517	0.461	0.498	0.442	0.449	0.372	0.426	0.353
	0.558	0.888	0.539	0.879	0.502	0.912	0.472	0.896
0.90	0.579	0.524	0.558	0.502	0.519	0.435	0.481	0.412
	0.617	0.912	0.596	0.903	0.574	0.942	0.544	0.928
0.95	0.686	0.637	0.663	0.612	0.661	0.591	0.632	0.546
	0.716	0.944	0.695	0.937	0.703	0.968	0.668	0.965

The posterior proportion, $P(\sum I_j / 570 | data)$, is calculated as the posterior (median) estimate of proportion of indicators being one among 570 pairs. The observed proportion is calculated as a proportion of non-zero posterior (median) estimate of $\beta_j = I_j \times \mu_j$ among 570 pairs.

lead to large deviation from the prior so that the posterior/observed proportion of eQTLs is always much smaller than the prior proportion. This means that data strongly support the conclusion of the absence of the regulation for a large number of pairs. In contrast, the analyses with the transformed data (by a gene-specific scaling factor) lead to the estimated proportions extremely close to the prior, indicating the lack of information in the transformed data.

The estimated (posterior/observed) proportion of eQTL is slightly smaller for the model with the prior assumption $\mu_j \sim N_+(0,100)$ than for the model with $\mu_j \sim N_+(2,100)$, the exception being the case of $s^e = 0.80$ (Table 1). Because two priors are almost equal, this result can be explained by the fact that the model assuming $\mu_j \sim N_+(0,100)$ prefers to ‘find’ neutral (or extremely small effect) eQTLs rather than requiring them to be any larger. For this to be true, the information content of the data also need to be extremely low and/or supportive for small-effect eQTLs. Moreover, the observed proportion is always smaller than the posterior proportion which may indicate that it is more easy to have the posterior median of $\beta_j = I_j \times \mu_j$ equal to 0 for few j (to downweight the observed proportion) than to downweight the posterior median of $\sum I_j$ (influencing on the posterior proportion). Note also that the estimated (posterior/observed) proportion of eQTL is somewhat larger for the analysis of the transformed data (by a common scaling factor) than the analyses of the original data (Table 1). This may indicate better fit (perhaps even overfit) of the model to the data, because the scaled data perfectly correspond to the model assumption $\sigma_0^2 = 1$.

Model assessment: To assess the goodness-of-fit of the model, we want to assess how well one can predict values of the gene expressions based on the model and the posteriors of the parameters. These posterior

predictions $E_{i,j}^*$ are then compared to the observed data of each individual $E_{i,j}$ and one obtains the prediction error $PE_{i,j} = (E_{i,j}^* - E_{i,j})$ as a simple difference between the two. In general—for robust predictions—instead of using the best-case scenario (that is, to evaluate posterior predictive distribution only at a point estimate, for example the posterior mode or median), one should use the whole predictive posterior distribution (that is, include uncertainty of the whole posterior distribution) and thus utilize the Bayesian model averaging (West *et al.*, 2006). However, we are here more interested in checking the best-case scenario, that is, to sample a gene-expression value for each individual from its posterior predictive distribution $p(E^* | I, \mu, A, G, \sigma_0^2)$ conditionally on the genotype data and the posterior estimates of the parameters (I, μ, A, σ_0^2) using posterior medians of the continuous parameters. To handle missing genotype data, we again assumed a uniform prior $p(G)$. We can then calculate the mean and the variance of the prediction errors under the different models considered. Such summaries are presented in Figure 5. Because the observed gene-expression values contained some missing observations, the mean and the variance were calculated only over individuals with the observations.

It becomes evident that the analysis of the transformed data (with a gene-specific scaling factor) resulted in predictions where all the existing information is lost (Figure 5a). The same was true for the prediction error variance (picture not shown). Because the mean prediction errors from analyses of the original data with both the known and unknown σ_0^2 as well as of the transformed data (by using a common scaling factor) with the known σ_0^2 all resulted into very similar pictures with minor numerical differences in the mean prediction errors, only one of them is shown in Figure 5b. In Figures 5c, d, and e, one can see how the prediction error variance

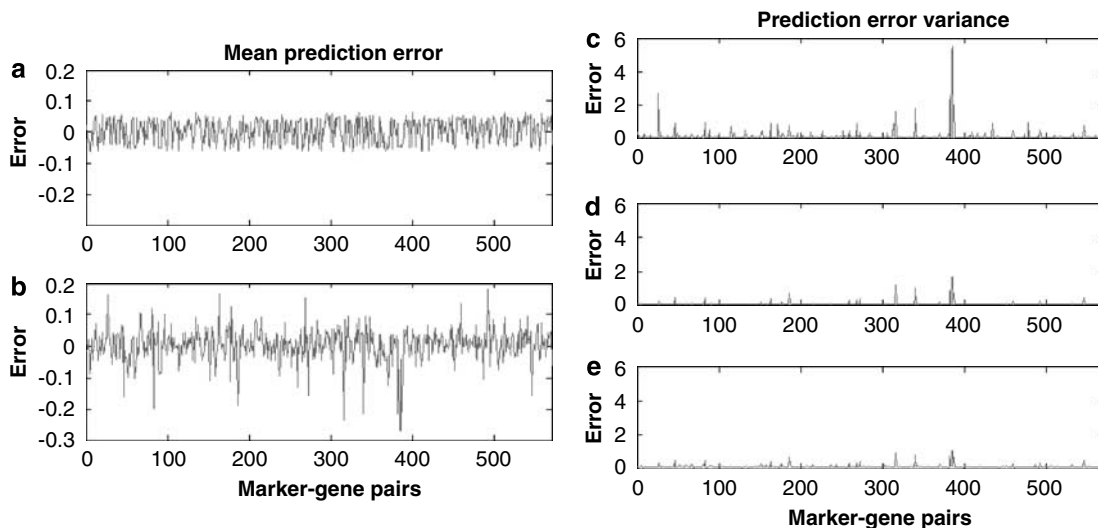


Figure 5 The mean and the variance of the individual-specific prediction error, which is a difference between the predicted and the observed gene expression of individual i at pair j . The quantities are calculated based on 1000 Markov Chain Monte Carlo (MCMC) samples from the posterior predictive distribution evaluated at the median (point estimate) of the posterior distribution of the model parameters from the model with $\mu \sim N_+(0,100)$. (a) The mean prediction error from the analysis of the transformed data (by using a gene-specific scaling factor) with the known $\sigma_0^2 = 1$. (b) The mean prediction error from the analysis of the original data with the known $\sigma_0^2 = 1$. (c) The prediction error variance from the analysis of the original data with unknown σ_0^2 . (d) The prediction error variance from the analysis of the transformed data (by using a common scaling factor) with the known $\sigma_0^2 = 1$.

decreases by treating σ_0^2 as an unknown variable or by using a data transformation and a common scaling factor. Even if the results suggest that the data transformation (scaling) is a reasonable way to proceed in this type of analysis, one should proceed with caution because some biological interactions may be lost or destroyed by the scaling (Jansen, 2003; Vormfelde and Brockmüller, 2007).

Simulated eQTL data with dependence between transcripts

Here we study prediction of the missing gene expressions based on the linked marker data and the observed values of the gene expressions on the flanking transcripts. We use the eQTL model with spatial dependencies here. Albeit this model may appear to be unrealistic, the results presented here arguably correspond to the more realistic case (for example, analysis of pathway membership model).

Simulating expressions: We use the same simulated marker data described above, but instead of utilizing all 102 markers we use only the 34 markers (evenly spaced at every 3 cM) on the first chromosome. Conditionally on the markers, we simulated 100 data sets with the correlated expression data, using the mean vector $\mathbf{\bar{O}} = (2, 2, \dots, 2)^T$, and the smoothing parameter values $\tau_0 = 4$ and $\lambda = 10$. This resulted in the average correlation of 0.7417 between any adjacent pair of effect sizes (μ_j, μ_{j+1}) in the data sets. Similarly, we also obtained the average correlation of 0.9722 by changing the smoothing parameter to $\lambda = 1$. In a following we refer to these two different generating models as 'a weak dependence ($\lambda = 10$)' model and 'a strong dependence ($\lambda = 1$)' model. All simulations were carried out in OpenBUGS using eQTL model with spatial dependencies.

Analysis of simulated eQTL data using the spatial dependence model for transcripts: Two simulated data sets were analyzed, a single realization generated with a weak dependence ($\lambda = 10$) model, and other obtained with 'a high dependence ($\lambda = 1$)' model. In the analysis stage, all the expression values at every other marker (in even numbers) in both data sets were coded as missing. These two data sets were both analyzed using two different eQTL models differing in the structure of the prior $p(\mu)$. The two eQTL models are the spatial dependence model (with the values $\mathbf{\bar{O}} = (0, 0, \dots, 0)^T$, $\tau_0 = 100$, and the known $\lambda = \{1, 10\}$) and the independence model (1) (with the uninformative prior $\mu_j \sim N(0, 100)$ in the positive range $[0, \infty[$, and the known values of $\sigma_0^2 = 1$ and $\alpha_0 = 0$).

In addition, to depict an interval of maximum estimation error, all the expressions were deleted from the first data set and analyzed using the independence model. The estimation using the spatial dependence model requires markedly more computational efforts because the model is more complicated. (Thus, for any practical settings, one should seriously consider some other computational tool than OpenBUGS.) All these analyses are summarized in Figure 6, except the independence model analysis of weakly correlated pair data with 100% of the expression measurements missing at every other marker (the picture is almost identical to d).

In general, the credible intervals (of the estimation error) of Figure 6 are constantly wide at every other marker in all cases because 100 % of the expressions were missing at those positions (that is, all the information comes through the dependence structure). However it becomes clear that the predictive properties of the spatial dependence model are slightly better than the independence model and its accuracy improves along with the increasing amount of dependence in the data (cf. the scales at the y -axis). On the other hand, the predictive accuracy of the independence model seems to stay

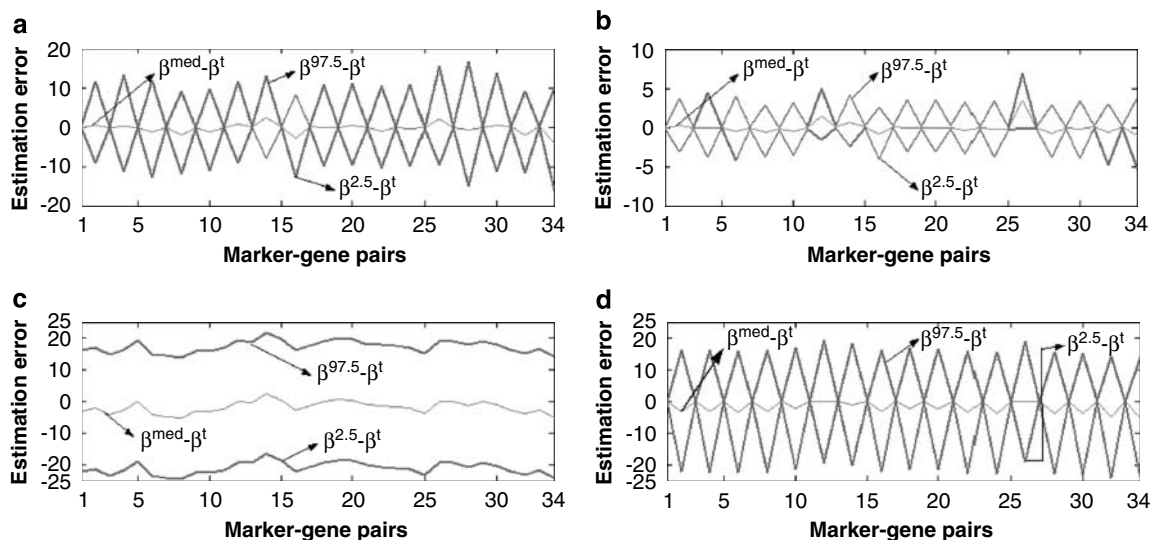


Figure 6 Summary of the estimation error of the posterior estimated regulatory effect β_j over 34 marker gene pairs. (a) The spatial dependence model analysis of weakly correlated pair data with 100% of the expression measurements missing at every other marker. (b) The spatial dependence model analysis of strongly correlated pair data with 100% of the expressions missing at every other marker. (c) The independence model analysis of weakly correlated pair data with all the expression measurements missing at every marker. (d) The independence model analysis of strongly correlated pair data with all the expression measurements missing at every other marker. The estimation error (of the posterior median and 2.5 and 97.5% quantiles) around the true simulated value β_j is presented at each position j . The pairs are shown on the x -axis and the estimation error on the y -axis.

practically constant while the amount of dependence in the data increases.

Simulated cQTL data

To test how well the cQTL model can handle the missing values among genotypes and expressions, we simulated eQTL data on backcross ($N = 200, N_p = 102$) as before (see details above) except that $\mu \sim N(0, 100)$ and $A_j = 1$. On the basis of the complete data and the cQTL model (2), nine components (1 marker, 5 expressions, and 3 genotype \times expression interactions) were used to generate the phenotypic values (that is, those components should exhibit non-zero cQTL effects with respect to the phenotype in the analysis). We used fixed value $\sigma_{\epsilon}^2 = 1/0.065 \approx 15.3846$ resulting to the joint heritability of the trait which was approximately 0.68. The actual effect sizes and types of the components are shown in Table 2. In the analysis stage, we again (in random locations) deleted 5% of the marker genotypes ($G_{i,j}$) and 50% of the gene expressions ($E_{i,j}$) from the complete data set. All phenotypes were assumed to be available and the genotypes and the expressions were assumed to be missing at random.

Analysis of simulated cQTL data: The data set was analyzed using two cQTL models, differing in the complexity of the missing data model for the missing values of expressions. The first model (*MD1*) is the one including the eQTL model (1) as missing data model, as presented in this article, and the second one (*MD2*) uses a much simpler model to handle the missing expressions, $E_{i,j} \sim N(0, \sigma_0^2)$, where $p(E|I, \mu, A, G, \sigma_0^2)$ is replaced simply by $p(E|\sigma_0^2)$. Note that the latter is close to the missing data model of Hoti and Sillanpää (2006), and it follows by assuming the additive polygenic (infinitely many loci) basis for gene expression. For the first model, the truncated normal prior: $\mu \sim N(0, 100)$ in the positive range $[0, \infty]$ is assumed and for the both of these two models, $\alpha_0 = 0$ and the prior $p(\sigma_0^2)$ is assumed to be an inverse Gamma (1,1) restricted to the range $[0.5, 10\,000]$. In both cases, 306 candidate terms (102 markers, 102 expressions, and 102 marker \times expression interactions) were considered in the model. For a Bernoulli parameter, we set $s^c = 0.0033 \approx 1/306$ which roughly corresponds to a single *a priori* associated component among the candidates. Note that because the phenotypes represent an outcome variable in the large cQTL model, the

posterior distributions of the eQTL model (1) parameters are now influenced by all (the missing and observed) expression values. This is contrary to modeling expressions as outcome variable in the plain eQTL model (see Equation (1) above). To estimate the parameters in *MD1* and *MD2*, the OpenBUGS 2.2.0 was ran for 110 000 MCMC iterations, with 10 000 burn-in. Surprisingly, we encountered a slight mixing problem in the sense that locations of the false positive cQTL signals, with small effect sizes, varied somewhat from one analysis to the next. However, there was only few such locations. It seemed that running longer chains did not influence to this property much. The convergence was inspected by comparing the results of different smaller runs. This was complicated by the fact that the running times for both analyses took more than a week on a personal computer. Unlike Hoti and Sillanpää (2006), we did not consider standardized effect sizes here. In Table 2, one can see the posterior weighted cQTL effects found for different genetic components under the two models. To define what is a cQTL, we had to choose rather high noise level (0.05 in analyses of Table 2). Among the markers, *MD1* found weak cQTL evidence for the correct locus (the pair 31) but it strongly supported also for the locus that is a false positive (the pair 15). For a comparison, all putative cQTL findings of *MD2* were false (the pairs 14, 15, 24, and 57), except a weak signal near the noise level (the pair 31). Because of the huge false positive signal at pair 14, we further checked the proportion of missing data at pair 14, which was unexpectedly less than an average ($\sim 4.5\%$ of the marker data and $\sim 47\%$ of the expression data). Among the expression effects, *MD1* correctly identified three out of five gene expressions (the pairs 14, 15 and 57) where, however, the cQTL evidence for the pair 57 was negligible. In addition, although negligible, also the cQTL was correctly estimated to have non-zero effect (-0.011) at position 100. Among the same candidates, *MD2* found only some negligible cQTL evidence (0.014) for the incorrect position (the pair 88, which was simulated to have an interaction effect). Finally, *MD1* correctly identified two out of three genotype \times expression interactions (the pairs 50 and 88) while *MD2* found none of them. It is good to emphasize here that although some marker gene pairs (14, 15 and 57) were interpreted as false positives among marker effects above, the same pairs originally had expression

Table 2 Posterior estimated (mean) and true cQTL effects under two models (*MD1* and *MD2*) for pairs where the true or estimated effect was nonnegligible (all values less than 0.05 are set to 0 or not shown)

Effects	Pair j									
	12	14	15	24	31	50	57	69	88	100
η_j^M true	0	0	0	0	-2.065	0	0	0	0	0
η_j^M MD1	0	0	1.089	0	-0.232	0	0.068	0	0	0
η_j^M MD2	0	-26.87	1.249	0.319	-0.070	0	0.538	0	0	0
η_j^E true	0	1.020	0.318	0	0	0	0.988	0.238	0	-0.976
η_j^E MD1	0	1.040	0.150	0	0	0	0.081	0	0	0
η_j^E MD2	0	0	0	0	0	0	0	0	0	0
η_j^{ME} true	1.021	0	0	0	0	4.056	0	0	-2.817	0
η_j^{ME} MD1	0	0	0	0	0	5.009	0	0	-3.940	0
η_j^{ME} MD2	0	0	0	0	0	0	0	0	0	0

The cQTL effects, at pair j , are shown for marker genotypes ($\eta_j^M = I_j^M \times \beta_j^M$), gene expressions ($\eta_j^E = I_j^E \times \beta_j^E$) and genotype \times expression interactions ($\eta_j^{ME} = I_j^{ME} \times \beta_j^{ME}$). The correctly identified pairs are highlighted in bold.

effects. Thus, they actually are false positives only in the sense of their effect types rather than their positions. To further experiment with *MD1*, we analyzed the same data with different missing data pattern (5% of the marker data and 30% of the expressions missing at random). This data was ran for 110 000 MCMC rounds with 10 000 burn-in. The results were quite similar to the ones of *MD1* in Table 2, except the huge marker effect (-26.46) and no expression effect at pair 14, similar to *MD2* above (results not shown). As a conclusion, it becomes clear that the more complicated model, *MD1*, outperforms the simpler model, *MD2*, and leads to better identification of cQTLs. Actually the poor performance of *MD2* indicates that the amount of missingness is very large, and it is helpful in such cases to utilize marker and phenotype information jointly to predict the missing values of expressions.

Discussion

We have presented here a new method for simultaneously estimating *cis*- and *trans*-acting eQTL effects as well as the cQTL effects among the preselected set of marker gene pairs. The method is based on hierarchical modeling so that the eQTL model is a part of the larger cQTL model. The both (eQTL and cQTL) models were tested as separate analyses in presence of missing data by assuming missing at random (Rubin, 1976). However, there is one important difference in these two analyses that needs more attention. Namely, in the plain eQTL analysis, the posterior distributions of the eQTL model parameters are influenced only by the observed part of the expression data, whereas in cQTL analysis imputed expression values also influence the posterior. Therefore, in cQTL analysis, one should be careful that the amount of missing data does not become larger than the observed part of the data (cf. Kilpikari and Sillanpää, 2003). Even if not detected here, there may still be some unwanted biases present in the estimates when the amount of missing data exceeds 50%. The presented method is, to our knowledge, the first attempt to model these two tasks simultaneously within a single modeling framework. Therefore, we want to here briefly discuss different future directions that we feel are central in this context.

Multiple trait analysis

In this article we have considered only a single phenotype at a time. However, using several traits simultaneously would be interesting extension to be considered in the future that definitely can provide more information on locating the cQTLs and eQTLs as well as on separating pleiotropy from close linkage. In case of pleiotropy, we can further consider separating marker effects from expression effects at same location. In addition, an interesting issue here is the comorbidity, association of two or more traits, which would provide insight on direct and indirect genetic effects (Smoller *et al.*, 2000; Robins *et al.*, 2001; Corander and Sillanpää, 2002; Grünwald, 2004; D Remington, North Carolina, personal communication). To study this issue, Li *et al.* (2006) presented the structural equation model, where hierarchical regression relationships between variables are determined. Verzilli *et al.* (2005) considered seemingly unrelated regressions model, where different sets

of single nucleotide polymorphisms can be taken as explanatory variables for each trait. As of their flexibility, these models could provide adequate framework for future extensions of our setup to multiple traits.

The central technical issue (for MCMC estimation and convergence of the Bayesian approach) due to the small sample size (number of individuals) is the efficient parametrization of the multiple trait model. The useful parametrization, in terms of restricting (between trait) effect ratio to be constant over alleles at each locus and gene correlations always to be either -1 and 1, has been proposed by Goddard (2001). For implementation, see Meuwissen and Goddard (2004). Also, an application of Bayesian variable selection for estimating non-zero elements of the covariance matrix has been suggested (Smith and Kohn, 2002).

Multiple gene models and model choice

The presented method is based on a single-eQTL model, which may limit an application of the method for eQTL mapping purposes, but it provides a new source of information for handling of missing gene expressions in the cQTL mapping context. In their real data application, Hoti and Sillanpää (2006) considered the eQTL model for a single expression phenotype, where an associated subset of multiple markers and expression levels (as well as their interaction components) were determined using Bayesian adaptive model selection. Perez-Enciso *et al.* (2007) proposed the use of support vector machines and stepwise regression approach for similar purpose. They also showed how the use of other expression levels as potential covariates in the model can improve the performance of eQTL mapping. Bhattacharjee and Sillanpää (2008) proposed the Bayesian cQTL model with the indicator variables (for model selection) to study stratified allele and expression effects to the phenotype using different clinical variables (for example, sex and onset) as stratifying factors. Recently, Jia and Xu (2007) presented a new Bayesian eQTL approach to simultaneously analyze hundreds of expression levels using a multiple marker model and model selection. Further studies are needed in this area, especially from the viewpoint of small sample size (number of individuals). This is because a small sample size has a direct impact on practical identifiability (multimodality of the posterior distribution) of the parameters and it largely determines what is a reasonable number of putative candidates and effects to be considered in the model (Hoti and Sillanpää, 2006). In such (sample size) assessment, one should also account for colinearity (correlateness) between candidates.

Pathway/dependence information

The pathway information is usually utilized to reduce the number of candidates in the cQTL analysis (Thomas, 2005), but we have presented here another way to incorporate dependence information between transcripts to the eQTL and cQTL analyses. If the eQTL model is omitted from the cQTL model, it is possible to model dependence between transcripts directly in their values of gene expressions. This represents again an alternative formulation of missing data model for expressions in the cQTL model context. Other than the missing data model, the dependence between candidates (markers/gene

expressions) of the cQTL model can be modeled also indirectly by introducing dependence prior for the model selection indicators (Sillanpää and Bhattacharjee, 2005).

See also the approach of Malo *et al.* (2008). To model dependence due to pathway membership in the cQTL analysis, Hung *et al.* (2004) have suggested the approach where the effects, of the markers (genes) being members of the same pathway, are exchangeable and arise from a common distribution. To consider more about pathway-based approaches, see Wang *et al.* (2007), and Luan and Li (2008).

The model specification code (written in OpenBUGS) is freely available for research purposes at <http://www.rni.helsinki.fi/~mjs/>.

Acknowledgements

We are grateful to Rashi Gupta for discussions, Andrew Thomas and one anonymous referee for their constructive comments on the article and Fabian Hoti for his help in the preliminary analysis of the real data. This work was supported by research grant no. 202324 from the Academy of Finland.

References

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.

Bhattacharjee M, Sillanpää MJ (2008). Bayesian joint disease-marker-expression analysis applied to clinical characteristics of chronic fatigue syndrome. (To appear in a book concluding the selected papers from CAMDA 2006.).

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

Bueno Filho JSS, Gilmour SG, Rosa GJS (2006). Design of microarray experiments for genetical genomics studies. *Genetics* **174**: 945–957.

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

Chen M, Kendziorski C (2007). A statistical framework for expression quantitative trait loci mapping. *Genetics* **177**: 761–771.

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

Conti DV, Witte JS (2003). Hierarchical modeling of linkage disequilibrium: genetic structure and spatial relations. *Am J Hum Genet* **72**: 351–363.

Corander J, Sillanpää MJ (2002). A unified approach to joint modeling of multiple quantitative and qualitative traits in gene mapping. *J Theor Biol* **218**: 435–446.

de Koning D-J, Haley CS (2005). Genetical genomics in humans and model organisms. *Trends Genet* **21**: 377–381.

Draghici S, Khatri P, Eklund AC, Swallasi Z (2006). Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet* **22**: 101–109.

Foss EJ, Radulovic D, Shaffer SA, Ruderfer DM, Bedalov A, Goodlett DR *et al.* (2007). Genetic basis of proteome variation in yeast. *Nat Genet* **39**: 1369–1375.

Fu J, Jansen RC (2006). Optimal design and analysis of genetic studies on gene expression. *Genetics* **172**: 1993–1999.

Gelfond JAL, Ibrahim JG, Zou F (2007). Proximity model for expression quantitative trait loci (eQTL) detection. *Biometrics* **63**: 1108–1116.

Gelman A (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1**: 515–533.

Geweke J (1996). Variable selection and model comparison in regression. In: Bernardo JM, Berger JO, Dawid AP, Smith AFM (eds). *Bayesian Statistics 5*. Oxford Press: Oxford, pp 609–620.

Gibson G, Weir B (2005). The quantitative genetics of transcription. *Trends Genet* **11**: 616–622.

Goddard ME (2001). The validity of genetic models underlying quantitative traits. *Livest Prod Sci* **72**: 117–127.

Grünwald M (2004). Genetic association studies with complex ascertainment. Licentiate thesis. Research Report 2004: 5. Mathematical Statistics, Stockholm University, Sweden.

Hazelton M, Gurrin LC (2003). A note on genetic variance components in mixed models. *Genet Epidemiol* **24**: 297–301.

Hoti F, Sillanpää MJ (2006). Bayesian mapping of genotype x expression interactions in quantitative and qualitative traits. *Heredity* **97**: 4–18.

Hung RJ, Brennan P, Malaveille C, Porru S, Donato F, Boffetta P *et al.* (2004). Using hierarchical modeling in genetic association studies with multiple markers: application to a case-control study of bladder cancer. *Cancer Epidemiol Biomarkers Prev* **13**: 1013–1021.

Jannink J-L (2005). Selective phenotyping to accurately map quantitative trait loci. *Crop Sci* **45**: 901–908.

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.

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

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

Jin C, Lan H, Attie AD, Churchill GA, Bulutuglo D, Yandell BS (2004). Selective phenotyping for increased efficiency in genetic mapping studies. *Genetics* **168**: 2285–2293.

Kendziorski CM, Chen M, Yuan M, Lan H, Attie AD (2006). Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics* **62**: 19–27.

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

Kuo L, Mallick B (1998). Variable selection for regression models. *Sankhya Ser B* **60**: 65–81.

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.

Li R, Tsaih S-W, Shockley K, Stylianou IM, Wergedal J, Paigen B *et al.* (2006). Structural model analysis of multiple quantitative traits. *PLoS Genet* **7**: e114.

Luan Y, Li H (2008). Group additive regression models for genomic data analysis. *Biostatistics* **9**: 100–113.

Malo N, Libiger O, Schork NJ (2008). Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. *Am J Hum Genet* **82**: 375–385.

Mehrabian M, Allayee H, Stockton J, Lum PY, Drake TA, Castellani LW *et al.* (2005). Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits. *Nat Genet* **37**: 1224–1233.

Meuwissen THE, Goddard ME (2004). Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. *Genet Sel Evol* **36**: 261–279.

Meuwissen THE, Hayes BJ, Goddard ME (2001). Prediction of total genetic value using genome-wide dense marker maps. *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.
- Mueller M, Goel A, Thimma M, Dickens NJ, Aitman TJ, Mangion J (2006). eQTL Explorer: integrated mining of combined linkage and expression experiments. *Bioinformatics* **22**: 509–511.
- Parmigiani G, Garrett ES, Irizarry R, Zeger SL (2003). *The Analysis of Gene Expression Data: Methods and Software*. Springer Verlag: New York.
- Perez-Enciso M, Quevedo JR, Bahamonde A (2007). Genetical genomics: use all data. *BMC Genomics* **8**: 69.
- 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.
- Qu Y, 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.
- Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I *et al.* (2000). Genome-wide location and function of DNA binding proteins. *Science* **290**: 2306–2309.
- Robins JM, Smoller JW, Lunetta KL (2001). On the validity of the TDT test in the presence of comorbidity and ascertainment bias. *Genet Epidemiol* **21**: 326–336.
- 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.
- Rubin DB (1976). Inference and missing data. *Biometrika* **63**: 581–592.
- Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D *et al.* (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet* **37**: 710–717.
- Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V *et al.* (2003). The genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**: 297–302.
- Sillanpää MJ, Arjas E (1998). Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics* **148**: 1373–1388.
- Sillanpää MJ, Bhattacharjee M (2005). Bayesian association-based fine mapping in small chromosomal segments. *Genetics* **169**: 427–439.
- Sillanpää MJ, Bhattacharjee M (2006). Association mapping of complex trait loci with context-dependent effects and unknown context variable. *Genetics* **174**: 1597–1611.
- Sladek R, Hudson TJ (2006). Elucidating *cis*- and *trans*-regulatory variation using genetical genomics. *Trends Genet* **22**: 245–250.
- Smith M, Kohn R (2002). Parsimonious covariance matrix estimation for longitudinal data. *J Am Stat Assoc* **97**: 1141–1153.
- Smoller J, Lunetta K, Robins J (2000). Implications of comorbidity and ascertainment bias for identifying disease genes. *Am J Med Genet* **96**: 817–822.
- Spiegelhalter D, Thomas A, Best N, Lunn D (2005). *WinBUGS User Manual, Version 2.10*. MRC Biostatistics Unit, Institute of Public Health: Cambridge, UK.
- Thomas A, O'Hara RB, Ligges U, Sturtz S (2006). Making BUGS open. *R News* **6**: 17–21.
- Thomas DC (2005). The need for a systematic approach to complex pathways in molecular epidemiology. *Cancer Epidemiol Biomarkers Prev* **14**: 557–559.
- Uimari P, Hoeschele I (1997). Mapping linked quantitative trait loci using Bayesian analysis and Markov chain Monte Carlo algorithms. *Genetics* **146**: 735–743.
- Uimari P, Thaller G, Hoeschele I (1996). The use of multiple markers in a Bayesian method for mapping quantitative trait loci. *Genetics* **143**: 1831–1842.
- Van Dongen S (2006). Prior specification in Bayesian statistics: three cautionary tales. *J Theor Biol* **242**: 90–100.
- Verzilli CJ, Stallard N, Whittaker JC (2005). Bayesian modeling of multivariate quantitative traits using seemingly unrelated regressions. *Genet Epidemiol* **28**: 313–325.
- Vormfelde SV, Brockmüller J (2007). On the value of haplotype-based genotype–phenotype analysis and on data transformation in pharmacogenetics and -genomics. *Nat Rev Genet* (01 Dec 2007).
- Wang K, Li M, Bucan M (2007). Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet* **81**: 1278–1283.
- Wang S, Basten CJ, Zeng Z-B (2006). *Windows QTL Cartographer 2.5*. Department of Statistics, North Carolina State University: Raleigh, NC.
- West M, Ginsburg GS, Huang AT, Nevin JR (2006). Embracing the complexity of genomic data for personalized medicine. *Genome Res* **16**: 559–566.
- Xu Z, Zou F, Vision TJ (2005). Improving quantitative trait loci mapping resolution in experimental crosses by the use of genotypically selected samples. *Genetics* **170**: 401–408.
- Yanai I, Korbel JO, Boue S, McWeeney SK, Bork P, Lercher MJ (2006). Similar gene expression profiles do not imply similar tissue function. *Trends Genet* **22**: 132–138.
- Yi N, George V, Allison DB (2003). Stochastic search variable selection for identifying multiple quantitative trait loci. *Genetics* **164**: 1129–1138.

Appendix: eQTL and cQTL models for F_2

Here we consider F_2 offspring resulting from a cross between two inbred lines, in which case the genotype value $G_{i,j}$ of individual i at marker j is 1 for genotype AA , -1 for genotype aa , and 0 for Aa . By assuming that only single marker (a major gene) is controlling each expression phenotype, there will be two equal sized (small) modes and one large mode in the tri-modal gene-expression frequency distribution following genotypic offspring ratio, which is 1:2:1 for the genotypes AA , Aa and aa , respectively. We assume that the mean of two homozygotes $\alpha_j = 0$ is located between the two smaller modes and μ_j is an additive effect (a deviation from the mean). In case $\mu_j = 0$ both distributions (modes) of AA and aa genotypes coincide. For F_2 , we need also the dominance parameter D_j to describe the effect of a heterozygote genotype Aa at each pair j . How far (and in which side) the large mode is from the mean $\alpha_j = 0$, is depending on a magnitude (and a sign) of the dominance D_j . In case $D_j = \mu_j = 0$ the distributions of all genotypes coincide and cannot be distinguished. Now, given the model parameters $(I_j, \mu_j, D_j, A_j, G_{i,j}, \alpha_j, \sigma_0^2)$, we can assume a following linear eQTL model $E_{i,j} = \alpha_j + I_j(\mu_j A_j G_{i,j} + D_j(1 - |G_{i,j}|)) + \varepsilon_{i,j}$. Again the residuals $\varepsilon_{i,j}$ (over pairs) are normally distributed with mean 0 and variance σ_0^2 , and the model selection indicator I_j is defined as before. The assignment variable A_j of pair j defines the sign of the additive effect, corresponding to the two orderings of the genotypes AA and aa . Note that, like in backcross, the dependence between transcripts is modeled only in the additive effects μ_j .

In case of F_2 intercross, the cQTL model is similar to before except that, $\beta_{j,1}^M = \beta_{j,1}^M$, $\beta_{j,2}^M$ and $\beta_{j,1}^{ME} = \beta_{j,1}^{ME}$, $\beta_{j,2}^{ME}$ are vectors containing two elements. In each case, the two elements are assumed to be exchangeable and arise from a common distribution.