frontiers in
GENETICS

ORIGINAL RESEARCH ARTICLE
published: 06 June 2012
doi: 10.3389/fgene.2012.00097

A hierarchical Bayesian approach to multi-trait clinical quantitative trait locus modeling

Crispin M. Mutshinda[1†], Neli Noykova[1] and Mikko J. Sillanpää[1,2,3,4]*

[1] *Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland*
[2] *Department of Agricultural Sciences, University of Helsinki, Helsinki, Finland*
[3] *Department of Mathematical Sciences, University of Oulu, Oulu, Finland*
[4] *Department of Biology, University of Oulu, Oulu, Finland*

Edited by:
Kenneth S. Kompass, University of California San Francisco, USA

Reviewed by:
Jian Li, Tulane University, USA
Bjarni V. Halldorsson, Reykjavik University, Iceland

***Correspondence:**
Mikko J. Sillanpää, Department of Mathematical Sciences, PO Box 3000, University of Oulu, FIN-90014 Oulu, Finland.
e-mail: mjs@rolf.helsinki.fi

[†]***Current address:***
Crispin M. Mutshinda, Department of Mathematics and Computer Science, Mount Allison University, York Street 67, Sackville, NB, Canada E4L 1E6.

Recent advances in high-throughput genotyping and transcript profiling technologies have enabled the inexpensive production of genome-wide dense marker maps in tandem with huge amounts of expression profiles. These large-scale data encompass valuable information about the genetic architecture of important phenotypic traits. Comprehensive models that combine molecular markers and gene transcript levels are increasingly advocated as an effective approach to dissecting the genetic architecture of complex phenotypic traits. The simultaneous utilization of marker and gene expression data to explain the variation in clinical quantitative trait, known as clinical quantitative trait locus (cQTL) mapping, poses challenges that are both conceptual and computational. Nonetheless, the hierarchical Bayesian (HB) modeling approach, in combination with modern computational tools such as Markov chain Monte Carlo (MCMC) simulation techniques, provides much versatility for cQTL analysis. Sillanpää and Noykova (2008) developed a HB model for single-trait cQTL analysis in inbred line cross-data using molecular markers, gene expressions, and marker-gene expression pairs. However, clinical traits generally relate to one another through environmental correlations and/or pleiotropy. A multi-trait approach can improve on the power to detect genetic effects and on their estimation precision. A multi-trait model also provides a framework for examining a number of biologically interesting hypotheses. In this paper we extend the HB cQTL model for inbred line crosses proposed by Sillanpää and Noykova to a multi-trait setting. We illustrate the implementation of our new model with simulated data, and evaluate the multi-trait model performance with regard to its single-trait counterpart. The data simulation process was based on the multi-trait cQTL model, assuming three traits with uncorrelated and correlated cQTL residuals, with the simulated data under uncorrelated cQTL residuals serving as our test set for comparing the performances of the multi-trait and single-trait models. The simulated data under correlated cQTL residuals were essentially used to assess how well our new model can estimate the cQTL residual covariance structure. The model fitting to the data was carried out by MCMC simulation through OpenBUGS. The multi-trait model outperformed its single-trait counterpart in identifying cQTLs, with a consistently lower false discovery rate. Moreover, the covariance matrix of cQTL residuals was typically estimated to an appreciable degree of precision under the multi-trait cQTL model, making our new model a promising approach to addressing a wide range of issues facing the analysis of correlated clinical traits.

Keywords: Bayesian multilevel modeling, genetic architecture, linked marker-expression pairs, pleiotropy

INTRODUCTION

Integrating genetic polymorphism and gene expression data to elucidate the genetic architecture and regulatory networks of complex clinical traits is a rousing trend in modern biology. This tendency owes much to the now established view (e.g., Schadt et al., 2005; Kendziorski et al., 2006; Lee et al., 2009; Mackay, 2009) that gene expression profiles usually act as intermediate phenotypes between genetic polymorphism and the phenotypic traits of interest.

The genomic loci associated with the variation in gene transcript levels, known as expression quantitative trait loci (eQTLs), can be identified through a standard quantitative trait locus (QTL) mapping framework, with transcript levels acting as surrogate for classical quantitative traits (Jansen and Nap, 2001; Schadt et al., 2003; Cheung et al., 2005; Drake et al., 2006; Breitling et al., 2008). An expression profile that is treated as a continuous trait for mapping purposes is called an expression trait (eTrait; Zou et al., 2007), and the genome-wide genetic analysis of gene expression data is known as genetical genomics (Jansen and Nap, 2001) or transcriptome mapping (Li and Deng, 2010).

An eQTL is said to be *cis*- or *trans*-acting (Brem et al., 2002), depending on its location with regard to the chromosomal position

of its target gene (i.e., the gene whose expression it regulates). A *cis* eQTL encompasses the genomic location of its target gene, whereas a *trans* eQTL maps to a distant genomic location. *Trans* eQTLs may aggregate in small segments of DNA sequences called genomic "hotspots" in which each eQTL may regulate a large number of gene transcripts (Breitling et al., 2008; Wu et al., 2008). It is, however, not straightforward to determine whether an eQTL acts *in cis* or *in trans*. One way out is to consider as *cis*-acting all eQTLs lying within a specific distance of their target genes, and view the ones that are far removed from their target genes as *trans*-acting (e.g., Brem et al., 2002; Wittkopp, 2005). Along these lines, Brem et al. (2002) used 10 kb as the threshold distance for distinguishing between *cis*- and *trans*-regulatory effects.

With the advent of high-throughput genotyping and transcript profiling technologies, it is now easy and inexpensive to concurrently generate genome-wide dense marker maps and huge amounts of expression profiles for each individual in a study population (Borevitz et al., 2003; Ronald et al., 2005). These large-scale data are generally littered with valuable information on the link between genetic polymorphisms and clinical traits of interest, and on the subtle molecular networks or pathways involved. The simultaneous utilization of marker and expression data to explain the variation in clinical quantitative traits is termed clinical quantitative trait locus (cQTL) analysis (Hoti and Sillanpää, 2006; Sillanpää and Noykova, 2008; Pikkuhookana and Sillanpää, 2009). cQTL analysis poses many problems and challenges, four of which are pointed out below.

(1) The high model dimensionality implied by the huge number of parameters undermines the effectiveness of standard statistical methods. (2) High correlations between predictors (markers or expressions) tend to reduce statistical power in the sense that, if one predictor shows a spurious association, its correlates will most likely show that same erroneous association. (3) The statistical issue of inflated false discovery rate (FDR) or type I error due to multiple testing (Kendziorski et al., 2006) limits the usefulness of single-locus testing procedures. A multi-locus approach provides more power for identifying the few potentially relevant loci to the phenotype-to-genotype association in both QTL and eQTL analyses. (4) Small sample size in terms of the number of individuals (de Koning and Haley, 2005) remains a problem in both QTL and eQTL analyses as the curse of dimensionality associated with the so-called "large p small n" problem is ever more ubiquitous. In this regard, regularization or shrinkage methods (e.g., Xu, 2003; Mutshinda and Sillanpää, 2010, 2011) are increasingly advocated as an effective way of reducing the model dimensionality in a regression set-up, by shrinking the effects of irrelevant covariates toward zero.

Hierarchical Bayesian (HB) modeling or Bayesian multilevel modeling (Gelman et al., 2003) provides a convenient approach for combining information from various data sources and accommodating uncertainty at different levels. By HB model, we understand a Bayesian model conceptualized in a hierarchical form in the sense that, the parameters involved in the likelihood function have priors, the parameters of which may also have priors involving a set of parameters, which may in turn have priors and so on, with the process coming to an end when no new priors are introduced (e.g., Mutshinda et al., 2008). In many cases, the HB prior specification provides the flexibility to define more realistic priors intended to match the requirements of the data at hand, while taking into account existing knowledge and expert opinion. It also helps enhance parameter estimation by "borrowing strength" from data used to estimate related quantities. With recent advances in computer intensive sampling-based methods such as Markov chain Monte Carlo (MCMC) simulation techniques (Gilks et al., 1996), the computational hurdles that have long prevented the broad application of HB modeling are no longer an issue; Bayesian models of arbitrary complexity are now being developed and implemented across a broad spectrum of scientific disciplines.

Sillanpää and Noykova (2008) developed a HB model for single-trait cQTL analysis in inbred line cross-data, using molecular markers, gene expressions, and marker-gene expression pairs. Their approach involved an eQTL model as missing data model for the intermediate link between markers and transcript levels in the determination of clinical phenotypic traits. The intermediate eQTL model can provide valuable insights into gene networks and molecular mechanisms linking genes to the clinical traits of interest.

It has, however, been pointed out (e.g., Mackay, 2009) that phenotypic traits do not exist in isolation; they often relate to one another through environmental correlations and pleiotropy. Many authors, including Jiang and Zeng (1995) and Liu et al. (2007, 2008) have pointed up a number of advantages of a joint analysis of multiple correlated traits over their separate analyses. These include the improvement on the statistical power to detect QTLs and on the precision of parameter estimation. Moreover, a multi-trait model provides a formal framework for examining a number of biologically interesting hypotheses regarding the underpinnings of genetic correlations between different traits. This understanding is crucial in animal and plant breeding where, as pointed out by Jiang and Zeng (1995), one of the major goals is to break unfavorable linkage.

The aim of this study is to extend the HB cQTL model for inbred line crosses proposed by Sillanpää and Noykova (2008) to a multi-trait setting, to illustrate the implementation or our new model with simulated the data, and evaluate its performance, using the single-trait counterpart as benchmark for comparison.

MATERIALS AND METHODS

DESCRIPTION OF THE INPUT DATA

Keeping with Sillanpää and Noykova (2008), we restrict attention to inbred line crosses such as backcross or double haploid progeny with one of two possible genotypes at any locus. However, the model can straightforwardly be adapted for the F_2 inter-cross design as discussed in the appendix of Sillanpää and Noykova (2008).

The input data involve molecular markers, gene expressions, and a set of clinical phenotypic traits of interests from each sampled individual. In addition, some marker-expression associations are *a priori* suggested for inclusion in the model, and may concern *cis*- or *trans*-regulatory effects. Hoti and Sillanpää (2006) refer to the marker-gene expression pairs as linked data. The linked data may result from oligonucleotide array data (Borevitz et al., 2003;

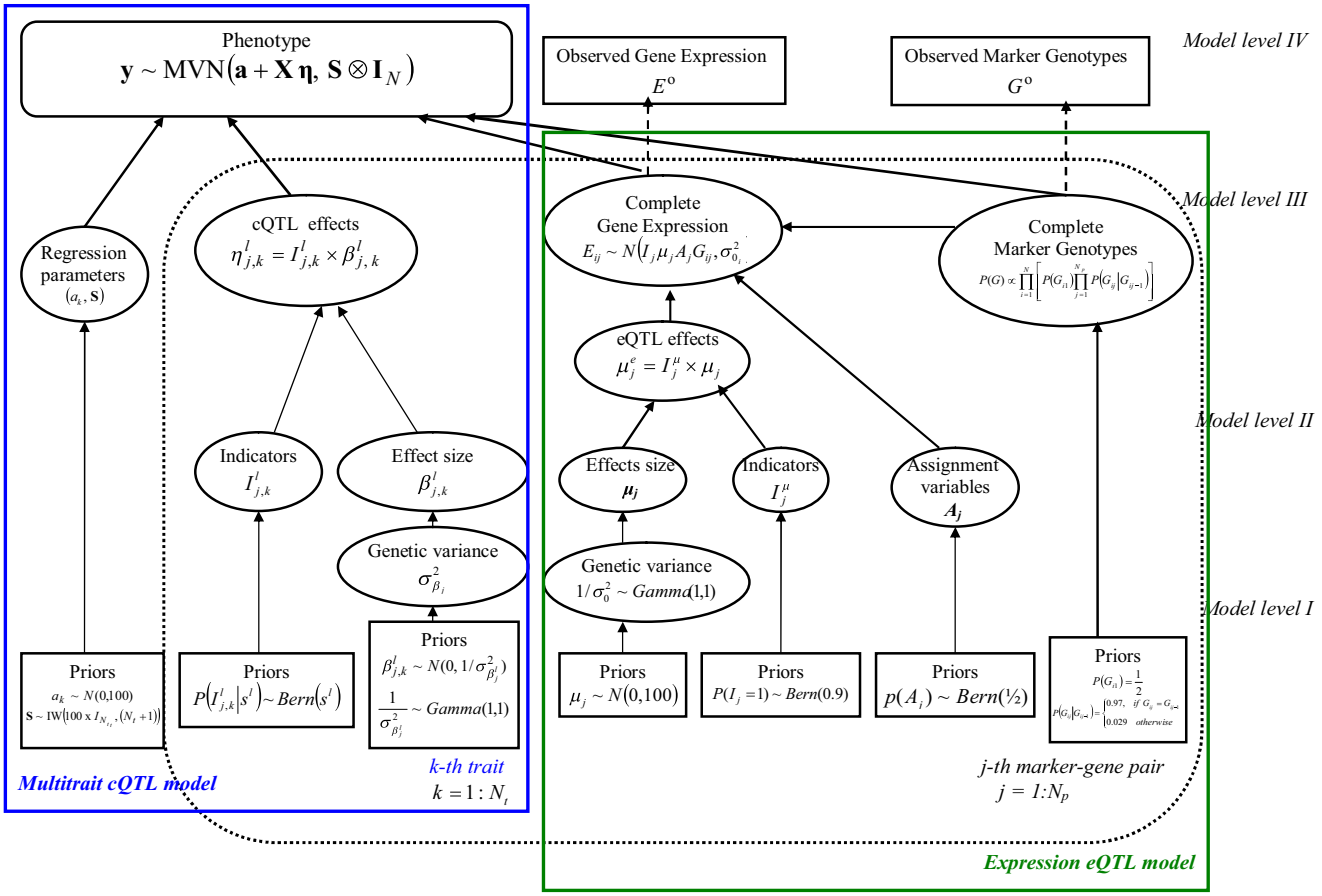

FIGURE 1 | Hierarchical directed acyclic graph (DAG) of the model structure. The pre-specified values (Model level I) or observed data (Model level IV) are given in boxes. Ellipses (Model levels II and III) refer to the unknown random variables, which are sampled. The solid arrows illustrate hierarchical dependencies, and the dotted arrows show deterministic dependences. The box given in bold indicates the multi-trait model structure.

Ronald et al., 2005) where markers and gene expression measurement are concurrently produced at every position, or be based on earlier findings of eQTL analyses or some known pathways. In cases where there is no *a priori* information to suggest the linked marker-expression pairs, these can be created from genetic distances, by assuming in *cis* effects between a marker and all genes falling within a specific genetic distance from it. As in Sillanpää and Noykova (2008), we assume each expression to be regulated by a single marker, without excluding the possibility for a marker to simultaneously regulate two or more expressions. In the latter case, the involved marker needs to be represented twice or as many times as required, the distance between its different copies being roughly zero.

SPECIFICATION OF THE MULTI-TRAIT cQTL MODEL

Let $\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{N_t}]$ denote the values of the N_t clinical quantitative traits of interest on the N study individuals, where $\mathbf{y}_k = (y_{k,1}, y_{k,2}, \ldots, y_{k,N})^{\mathrm{T}}$ represents the measurements of the kth trait ($k = 1 \ldots N_t$). For each trait, the cQTL model of Sillanpää and Noykova (2008) is assumed. That is,

$$\mathbf{y}_k = a_k \mathbf{1} + \tilde{\mathbf{X}} \boldsymbol{\eta}_k + \mathbf{e}_k, \quad (1)$$

where a_k is the population intercept for the kth trait, $\mathbf{1}$ is the $N \times 1$ vector of ones, $\mathbf{e}_k = (e_{k,1}, e_{k,2}, \ldots, e_{k,N})^{\mathrm{T}}$ is the residual vector associated with the kth trait, and $\tilde{\mathbf{X}}$ is the design matrix involving N_P markers ($\mathbf{G}$), N_P expressions ($\mathbf{E}$), and N_P marker-expression pairs ($\mathbf{GE}$) organized as $\tilde{\mathbf{X}} = [\mathbf{G} \mid \mathbf{E} \mid \mathbf{GE}]$. The parameter vector $\boldsymbol{\eta}_k$ therefore, describes the regulatory effect of genetic data on the kth trait. The full multi-trait cQTL model can be compactly written as

$$\mathbf{y} = \mathbf{a} + \mathbf{X}\boldsymbol{\eta} + \mathbf{e}, \quad (2)$$

where $\mathbf{a} = (a_1\mathbf{1}, a_2\mathbf{1}, \ldots, a_{N_t}\mathbf{1})$, $\mathbf{X}$ is a block-diagonal matrix comprising N_t blocks identical to $\tilde{\mathbf{X}}$, $\boldsymbol{\eta} = \mathbf{I} \bullet \boldsymbol{\beta}$ is the $3N_pN_t \times 1$ vector of regression coefficients to be estimated from the data, $\bullet$ denotes the entry-wise (Hadamard or Schur) product, $\mathbf{I}$ represents a $3N_pN_t \times 1$ vector of indicators, and $\boldsymbol{\beta}$ is the $3N_pN_t \times 1$ vector of genetic effects. As pointed out earlier, the method is developed for experimental crosses such as backcross or double haploid progeny with only one of two possible genotypes at any locus. Assuming that one genotype is coded as 1 and the other as -1, the size of the cQTL effects is represented by the quantity $2\mathbf{I}\boldsymbol{\beta}$.

The regression terms $\boldsymbol{\eta}^{\mathbf{M}}$, $\boldsymbol{\eta}^{\mathbf{E}}$, and $\boldsymbol{\eta}^{\mathbf{ME}}$ are respectively related to the marker genotypes, the expression measurements, and the mixed marker-expression pairs; $\mathbf{e} = (\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_{N_t})$ is the $N_t N \times 1$ residual vector assumed to follow a multivariate normal distribution $\mathbf{e} \sim MVN(\widetilde{\mathbf{0}}, \mathbf{S} \otimes \mathbf{I}_N)$ with the $N_t N \times 1$ vector $\widetilde{\mathbf{0}} = (0, 0, \ldots, 0)^T$ as mean and a $(N_t N \times N_t N)$ covariance matrix $\Sigma = \mathbf{S} \otimes \mathbf{I}_N$, where $\otimes$ denotes the Kronecker product operator. Here $\mathbf{S}$ is an $N_t \times N_t$ covariance matrix describing the variances and the (within individual) dependence between the residuals of different traits and $\mathbf{I}_N$ is the $N \times N$ identity matrix. This said, the distribution of $\mathbf{y}$ is given by $\mathbf{y} \sim MVN(\mathbf{a} + \mathbf{X}\boldsymbol{\eta}, \mathbf{S} \otimes \mathbf{I}_N)$, where $\mathbf{a}$, $\mathbf{X}$, and $\boldsymbol{\eta}$ are defined above.

HIERARCHICAL STRUCTURE OF THE MULTI-TRAIT cQTL MODEL

Our HB multi-trait cQTL model comprises four hierarchical levels as graphically depicted in **Figure 1**. Note that the intermediate eQTL model, presented as a shadowed box in the figure, is exactly the same as the eQTL part of the single-trait cQTL model (Sillanpää and Noykova, 2008). A detailed description of each hierarchical level is given below.

Model level IV

The highest level (level IV) of our HB model is represented by data vector $\mathbf{D} = (\mathbf{E}^0, \mathbf{G}^0, \mathbf{y})$.

Here the phenotypic data $\mathbf{y}$ (modeled by Eq. 1) for all N_t traits are assumed to be available with no missingness, while the observed gene expressions $\mathbf{E}^{\mathrm{o}}$ and marker genotypes $\mathbf{G}^{\mathrm{o}}$ may involve some missing values. The complete marker and expression data are respectively denoted by $\mathbf{G}$ and $\mathbf{E}$.

The parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}^{\mathbf{c}}, \boldsymbol{\theta}^{\mathbf{e}})$ to be estimated can be partitioned into two groups, namely $\boldsymbol{\theta}^{\mathbf{c}} = (\mathbf{I}, \boldsymbol{\beta}, \mathbf{a}, \mathbf{S}) = (\boldsymbol{\eta}, \mathbf{a}, \mathbf{S})$ which are directly involved in cQTL model (2), and $\boldsymbol{\theta}^{\mathbf{e}} = (\mathbf{I}^{\boldsymbol{\mu}}, \boldsymbol{\mu}, \mathbf{A}, \mathbf{G}, \mathbf{E}, \sigma_0^2)$ used in the intermediate eQTL model. The eQTL model parameter σ_0^2 is the expression variance, $\mathbf{I}^{\boldsymbol{\mu}}$ is a vector of indicators, $\boldsymbol{\mu}$ is the vector of eQTL effect sizes, and $\mathbf{A}$ comprises the assignment variables which, as in Sillanpää and Noykova (2008), define the expression eQTL regulatory effects of the marker-expression pairs.

Table 1 | Locations of the non-zero effects: for the first trait, there is marker $\eta^{M}_{24,1}$, and expression $\eta^{E}_{14,1}$ components, for the second trait, marker $\eta^{M}_{24,2}$, and a mixed genotype × expression interaction $\eta^{ME}_{4,2}$, and for the third trait, marker $\eta^{M}_{24,3}$, and expression $\eta^{E}_{18,3}$.

Location (j, k) of the regulatory effect $\eta_{j,k}$		Size of the effects $\eta^{l}_{j,k}$, $l = \{M, E, ME\}$		
Pair j	**Trait k**	η^{M}_{j}	η^{E}_{j}	η^{ME}_{j}
		True value	**True value**	**True value**
4	1	0	0	0
	2	0	0	**6**
	3	0	0	0
14	1	0	**6**	0
	2	0	0	0
	3	0	0	0
18	1	0	0	0
	2	0	0	0
	3	0	**6**	0
24	1	**6**	0	0
	2	**6**	0	0
	3	**6**	0	0

According the Bayes theorem $p(\boldsymbol{\theta} \mid \mathbf{D}) \propto p(\boldsymbol{D}, \boldsymbol{\theta}) = p(\boldsymbol{D} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$, which is equivalent to $p(\boldsymbol{\theta}^{\mathrm{c}}, \boldsymbol{\theta}^{\mathrm{e}} \mid \mathbf{E}^{\mathrm{O}}, \mathbf{G}^{\mathrm{O}}, \mathrm{y}) \propto p(\mathbf{I}, \boldsymbol{\beta}, \mathbf{a}, \mathbf{S}, \boldsymbol{\theta}^{\mathrm{e}}, \mathbf{E}^{\mathrm{o}}, \mathbf{G}^{\mathrm{o}}, \mathrm{y})$. This can be further factorized (according to the conditional independence assumptions made) to the form

$$p(\boldsymbol{\theta}^{c}, \boldsymbol{\theta}^{e} | \mathbf{E}^{o}, \mathbf{G}^{o}, \mathbf{y}) \propto p(\mathbf{y} | \mathbf{I}, \boldsymbol{\beta}, \mathbf{a}, \mathbf{S}, \mathbf{E}, \mathbf{G}) p(\mathbf{I} | \mathbf{s}) p(\boldsymbol{\beta} | \sigma \rho_{\boldsymbol{\beta}}^2) \times p(\sigma_{\boldsymbol{\beta}}^2) p(\mathbf{a}) p(\mathbf{S}) p(\mathbf{E}^{o}, \mathbf{G}^{o}, \boldsymbol{\theta}^{\mathbf{e}}). \quad (3)$$

The likelihood function associated with the multi-trait cQTL model (2) is given by

$$p\left(\mathbf{y} | \mathbf{I}, \boldsymbol{\beta}, \mathbf{a}, \mathbf{S}, \mathbf{E}, \mathbf{G}\right) = (2\pi)^{-N/2} |\mathbf{S} \otimes \mathbf{I}_N|^{-1/2} \exp\{-(\mathbf{y} - \mathbf{a} - \mathbf{X}\boldsymbol{\eta})^{\mathrm{T}} (\mathbf{S} \otimes \mathbf{I}_N)^{-1} (\mathbf{y} - \mathbf{a} - \mathbf{X}\boldsymbol{\eta}) / 2\}, \quad (4)$$

where $\boldsymbol{\eta} = \mathbf{I} \bullet \boldsymbol{\beta}$, and $|\mathbf{M}|$ denotes the determinant of $\mathbf{M}$. The other distributions on the right hand side (RHS) of Eq. 3 are described at lower hierarchical levels.

Model levels II and III

The intermediate hierarchical levels II and III involve models for the unknown parameters, as well as models for the complete marker genotype and gene expression data. The coefficients $\boldsymbol{\eta} = \mathbf{I} \bullet \boldsymbol{\beta}$ in the cQTL regression model (2) are formed on Model level III, whereas models for the genetic effects $\boldsymbol{\beta}$ and the eQTL effect sizes $\boldsymbol{\mu}$ appear on level II. The models of complete expression $\mathbf{E}$ and marker data $\mathbf{G}$ are given in model level III of eQTL model (Sillanpää and Noykova, 2008).

Model level I

The lowest hierarchical level (level I) consists of all pre-specified parameter and variable values.

At this level the rest of the terms on the RHS of Eq. 3 are defined. For the second term, we assume independence, so that $p(\mathbf{I}|\mathbf{s}) = \prod_{k=1}^{N_T} \prod_{j=1}^{N_p} p(I^{M}_{j,k}|s^{M}) p\,(I^{E}_{j,k}|s^{E}) p\,(I^{ME}_{j,k}|s^{ME})$, where $p(I^{l}_{j,k}|s^{l}) \sim Bernoulli(s^{l})$ is a Bernoulli-distributed indicator associated with the jth component of type $l = \{M, E, ME\}$ with regard to trait k. As for the single-trait cQTL model, we assume that $0 < s^{l} \leq 1/2$ is very small for all l components, implying a small probability that the corresponding candidate is associated with the trait.

For the third term, we assume conditional independence. That is, $p(\boldsymbol{\beta}|\boldsymbol{\sigma}_{\boldsymbol{\beta}}^2) = \prod_{k=1}^{N_T} \prod_{j=1}^{N_p} p\,(\beta^{M}_{j,k}|\sigma^2_{\beta^{M}_{j}}) p\,(\beta^{E}_{j,k}|\sigma^2_{\beta^{E}_{j}}) p\,(\beta^{ME}_{j,k}|\sigma^2_{\beta^{ME}_{j}})$. For each trait k, we assume that $\beta^{l}_{j,k}|\sigma^2_{\beta^{l}_{j}} \sim N(0, \sigma^2_{\beta^{l}_{j}})$, for $j = 1, \ldots, N_p$ and $l \in \{M, E, ME\}$.

For the genetic variances, we assume that $p(\boldsymbol{\sigma}_{\boldsymbol{\beta}}^2) = \prod_{k=1}^{N_T} \prod_{j=1}^{N_p} p(\sigma^2_{\beta^{M}_{j}}) p(\sigma^2_{\beta^{E}_{j}}) p\,(\sigma^2_{\beta^{ME}_{j}})$, and impose $InvGa(1, 1)$ priors

Table 2 | True and estimated (posterior means) cQTL effects under the MD1 version of the HB multi-trait cQTL model with 10% markers and 10% expressions coded as missing and different values of Bernoulli parameter s^l.

Location (j, k) of the effect $\eta_{j,k}$		Size of the regulatory effects $\eta^l_{j,k}$, $l = \{M, E, ME\}$								
Pair j	Trait k	η^M_j			η^E_j			η^{ME}_j		
		True value	Estimated		True value	Estimated		True value	Estimated	
			$s^l = 0.013$	$s^l = 0.09$		$s^l = 0.013$	$s^l = 0.09$		$s^l = 0.013$	$s^l = 0.09$
4	1	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	**6**	**4.01**	**4.11**
	3	0	0	0	0	0	0	0	0	0
14	1	0	0	0	**6**	**5.52**	**5.53**	0	0	0
	2	0	0	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0	0
16	1	0	0	0	0	0	0	0	0	0
	2	0	0	**−0.27**	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0	0
18	1	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0	0
	3	0	0	0	**6**	**5.95**	**5.95**	0	0	0
22	1	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	**0.65**	**0.63**
	3	0	0	0	0	0	0	0		0
24	1	**6**	**5.05**	**5.20**	0	0	0	0	0	0
	2	**6**	**5.57**	**5.68**	0	0	0	0	0	0
	3	**6**	**5.25**	**5.36**	0	0	0	0	0	0

A bold font is used to indicate positions where the true or estimated effect was non-zero ($|\bar{\eta}^l_{j,k}| \geqslant 0.2$). The shaded cells indicate false positives and false negatives.

independently on the variance parameters $\sigma^2_{\beta^l_j}$ for $j = 1, \ldots, N_p$. That is, $1/\sigma^2_{\beta^l_j} \sim Ga(1, 1)$, where $Ga(\alpha, \beta)$ denote the Gamma distribution with mean α/β and variance α/β^2. In addition, we impose the following (non-informative) normal prior distribution on the parameters $a_k \sim N(0, 100)$.

We place on the $N_t \times N_t$ residual covariance matrix $\mathbf{S}$ an inverse Wishart prior with matrix parameter (or prior covariance matrix) $100 \times \mathbf{I}_{N_t}$ and N_t degrees of freedom, or equivalently, a Wishart prior with matrix parameter $100 \times \mathbf{I}_{N_t}$ and N_t degrees of freedom on the precision matrix $\mathbf{S}^{-1}$. That is, $p(\mathbf{S}^{-1}) \propto |\mathbf{S}^{-1}|^{\frac{1}{2}[(N_t)-k-1]} \exp\{-\frac{1}{2} trace[(100 \times \mathbf{I}_{N_t})^{-1} \mathbf{S}^{-1}]\}$. Note that the number of degrees of freedom is set to be the largest possible, i.e., the rank of $\mathbf{S}$, to convey a lack of prior information.

For the last term in the RHS of Eq. 3, the joint distribution $p(\mathbf{E}^o, \mathbf{G}^o, \boldsymbol{\theta}^e)$, is a part of eQTL model and is described in details in Sillanpää and Noykova (2008).

APPLICATIONS

Simulation of the multi-trait cQTL data

We used the same marker and expression data from backcross inbred line simulation experiment as Sillanpää and Noykova (2008), in order to compare the performances of the multi-trait and single-trait HB cQTL models. The expression data were simulated through the OpenBUGS 2.2.0 program (Thomas et al., 2006) conditionally on marker data, using the eQTL part of the multi-trait cQTL model. Because of the high computational burden of MCMC sampling, we chose a smaller subset from the complete simulated marker and expression data. We also reduced the population size from $N = 200$ to $N = 100$ individuals, and the number of marker-gene pairs from $N_p = 102$ to $N_p = 25$ so that the markers spanned only the first chromosome. The multi-trait clinical cQTL data were subsequently simulated using the already generated marker and expression data, through the multi-trait cQTL model (1), assuming a fairly small ($N_t = 3$) number of traits.

Following Sillanpää and Noykova (2008), we chose two non-zero regulatory effects $\boldsymbol{\eta}^l_{j,k}, l \in \{M, E, ME\}, j = 1, \ldots, N_p$, from every trait $k = 1, 2, 3$, to generate the phenotypic values for the multi-trait cQTL analysis. For the first trait we chose one marker $\boldsymbol{\eta}^M_{24,1}$, and one expression $\boldsymbol{\eta}^E_{14,1}$ components to be non-zero, for the second trait, one marker $\boldsymbol{\eta}^M_{24,2}$, and one mixed genotype × expression interaction $\boldsymbol{\eta}^{ME}_{4,2}$, and for the third trait, one marker $\boldsymbol{\eta}^M_{24,3}$, and one expression $\boldsymbol{\eta}^E_{18,3}$ components (**Table 1**). We fixed all non-zero effect sizes to the arbitrarily chosen value $\beta^M_{j,k} = \beta^E_{j,k} = \beta^{ME}_{j,k} = 6$. With two non-zero effects out of 25 candidates, the simulated value of $s^l = P(I^l_{j,k} = 1)$ is $2/25 = 0.08$ for all $l \in \{M, E, ME\}$.

We investigated the model performance in the presence of uncorrelated and correlated cQTL residuals, noting that uncorrelated cQTL residuals do not necessary imply uncorrelated traits since the traits can still be correlated under uncorrelated cQTL

Table 3 | True and estimated (posterior means) cQTL effects under the MD1 version of the HB multi-trait cQTL model with 10% markers and 30% expressions coded as missing and different values of Bernoulli parameter s^I.

Location (j, k) of the effect $\eta_{j,k}$		Size of the regulatory effects $\eta^I_{j,k}$, $I = \{M, E, ME\}$								
Pair j	Trait k	η^M_j			η^E_j			η^{ME}_j		
		True value	Estimated		True value	Estimated		True value	Estimated	
			$s^I = 0.013$	$s^I = 0.09$		$s^I = 0.013$	$s^I = 0.09$		$s^I = 0.013$	$s^I = 0.09$
2	1	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0	**0.26**
	3	0	0	0	0	0	0	0	0	0
4	1	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	**6**	**5.96**	**0**
	3	0	0	0	0	0	0	0	0	0
9	1	0	0	0	0	0	0	0	0	0
	2	0	0	**−12.87**	0	0	**−1.46**	0	0	**7.14**
	3	0	0	0	0	0	0	0	0	0
14	1	0	0	0	**6**	**5.78**	**5.52**	0	0	0
	2	0	0	**−0.2**	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0	0
18	1	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0	0
	3	0	0	0	**6**	**5.66**	**5.77**	0	0	0
21	1	0	0	0	0	0	0	0	0	0
	2	0	0	**0.26**	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0	0
24	1	**6**	**0.24**	**2.78**	0	0	0	0	0	0
	2	**6**	**1.09**	**3.40**	0	0	0	0	0	0
	3	**6**	**0.46**	**3.50**	0	0	0	0	0	0

A bold font is used to indicate positions where the true or estimated effect was non-zero ($|\bar{\eta}^I_{j,k}| \geq 0.2$). The shaded cells indicate false positives and false negatives.

residuals owing for instance to pleiotropy. The simulated data with uncorrelated cQTL residual represented our full test set for comparing the multi-trait and single-trait models, whereas the simulated data with correlated cQTL residuals data were merely used to test how well our multi-trait model is able to estimate the cQTL residual covariance structure.

To simulate data with small heritabilities and uncorrelated cQTL residuals, the elements of the residual covariance matrix $\mathbf{S}$ were arbitrarily fixed as $\mathbf{S} = \begin{pmatrix} 139.1 & & \\ 0 & 123.3 & \\ 0 & 0 & 128.1 \end{pmatrix}$, where the elements above the main diagonal have been omitted due to symmetry. On average, the standard deviations $\hat{\sigma}_{Y_k}$ of the simulated cQTL data over the $N = 100$ individuals for Trait 1, Trait 2, and Trait 3 were 13.62, 12.92, and 12.50, respectively, implying a joint heritability $\mathbf{h}^2 \approx (0.25, 0.26, 0.18)^{\mathrm{T}}$, where $h^2_k = (\hat{\sigma}^2_{Y_k} - \sigma^2_{S_k})/\hat{\sigma}^2_{Y_k}$, $k = 1, 2, 3$ and $\sigma^2_{S_k} = \mathbf{S}_{kk}$. Although these heritability values are small and may not provide the best conditions for investigating the model behavior, they do reflect the reality of values that are commonly encountered in real-world genetic data.

To simulate data with large heritabilities and correlated cQTL residuals, we set the elements of the residual covariance matrix as $\mathbf{S} = \begin{pmatrix} 154 & & \\ 130 & 112 & \\ 133 & 113 & 117 \end{pmatrix}$. On average, the standard deviations of the simulated cQTL data over the $N = 100$ individuals for Traits 1, 2, and 3 were 17.45, 17.92, and 16.75, respectively, implying a joint heritability $\mathbf{h}^2 \approx (0.49, 0.65, 0.58)^{\mathrm{T}}$.

Analysis of simulated data

Simulation of missing marker and expression data. Often considerable amount of missing marker and expression data may occur at random positions in the data matrix. Also, due to financial constraints, marker data may be available for much larger group of individuals than expression data. Following Sillanpää and Noykova (2008), we considered the following missing data scenarios for simulating data under uncorrelated cQTL residuals, in order to investigate the sensitivity of the method/model to the amount of randomly missing values: (1) 10% of both marker genotypes $G_{i,j}$ and gene expressions $E_{i,j}$ coded as missing. (2) 10% of marker genotypes $G_{i,j}$ and 30% of gene expressions $E_{i,j}$ coded as missing. We also investigated a third scenario with 10% of marker genotypes $G_{i,j}$ and 50% of gene expressions $E_{i,j}$ coded as missing, but 50% turned out to be a too high and

Table 4 | True and estimated (posterior means) cQTL effects under the MD2 version of the HB multi-trait cQTL model with 10% markers and 10% expressions coded as missing and different values of Bernoulli parameter s^l.

Location (j, k) of the effect $\eta_{j,k}$		Size of the regulatory effects $\eta^l_{j,k}$, $l = \{M, E, ME\}$								
Pair j	Trait k	η^M_j			η^E_j			η^{ME}_j		
		True value	Estimated		True value	Estimated		True value	Estimated	
			$s^l = 0.013$	$s^l = 0.09$		$s^l = 0.013$	$s^l = 0.09$		$s^l = 0.013$	$s^l = 0.09$
4	1	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	**6**	**0**	**0**
	3	0	0	0	0	0	0	0	0	0
14	1	0	0	0	**6**	**5.52**	**5.50**	0	0	0
	2	0	0	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0	0
18	1	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0	0
	3	0	0	0	**6**	**5.85**	**5.90**	0	0	0
22	1	0	0	0	0	0	0	0	0	0
	2	0	0	**0.22**	0	0	0	0	**1.91**	**1.91**
	3	0	0	0	0	0	0	0	0	0
24	1	**6**	**1.54**	**4.36**	0	0	0	0	0	0
	2	**6**	**0.95**	**2.82**	0	0	**−0.31**	0	0	0
	3	**6**	**2.03**	**4.59**	0	0	0	0	0	0

A bold font is used to indicate positions where the true or estimated effect was non-zero ($|\bar{\eta}^l_{j,k}| \geqslant 0.2$). The shaded cells indicate false positives and false negatives.

inconclusive amount of missingness. We do not report the results for this case.

Data analyses under the multi-trait and single-trait cQTL models. We analyzed the simulated (uncorrelated cQTL residual) data under missing data scenarios 1 and 2 using two different missing data models namely, the model MD1, shown on **Figure 1**, where $E_{i,j} \sim N(I_j \mu_j A_j G_{i,j}, \sigma^2_0)$ for each individual i and marker-gene pair j, and the simpler MD2 model $E_{i,j} \sim N(0, \sigma^2_0)$, where $p(E|I, \mu, A, G, \sigma^2_0)$ is simply replaced by $p(E|\sigma^2_0)$.

For both the MD1 and MD2 versions of the model, we assumed a Bernoulli prior for indicators $I^l_{k,j}|s^l \sim \text{Bern}(s^l)$, with two different pre-specified parameter values for s^l namely, (1) $s^l = 0.013 = 1/(3 \times 25)$, which implies fewer non-zero indicator elements than the true simulated value, and (2) $s^l = 0.09$ implying a slightly larger proportion of non-zero effects.

We first analyzed the simulated data using our multi-trait model, with s^l set to 0.013 and 0.09, and subsequently fitted the single-trait cQTL model of Sillanpää and Noykova (2008) to each trait separately, with s^l set to 0.0033, 0.013, and 0.09.

We used MCMC simulation through the Bayesian freeware OpenBUGS 2.2.0 (Thomas et al., 2006) to sample from the joint posterior of the model parameters. The BUGS code is available from the authors upon request. The reported results are based on 100,000 MCMC iterations, the first 10,000 of which were discarded as burn-in. The convergence of the MCMC was assessed through visual inspection of trace plots. The 100,000 MCMC iterations took roughly 256,000 and 59,000 s for the multi-trait and single-trait models, respectively on a PC equipped with an Intel(R) Core(TM)2 Duo CPU T550 at 1.83G Hz and 3.00GB of RAM.

We focus attention on the estimates (posterior means), $\bar{\eta}^l_{j,k}$, of the cQTL effects $\eta^l_{j,k}$, $l = \{M, E, ME\}$, $j = 1, \ldots, N_p$ and as a rule of thumb, we consider $\eta^l_{j,k}$ to be non-negligible if its posterior mean is equal or larger than 0.2 in absolute value i.e., if $|\bar{\eta}^l_{j,k}| \geq 0.2$. Thus, all estimated $\eta^l_{j,k}$ such that $|\bar{\eta}^l_{j,k}| < 0.2$ were set to zero and deemed negligible.

RESULTS

Under uncorrelated cQTL residual data, the multi-trait model broadly outperformed its single-trait counterpart. It is well known that the single-trait approach is prone to poor statistical power in the presence of correlated responses. In what follows, we provide a fairly detailed account of the results concerning the analysis based on simulated data under uncorrelated cQTL residuals, and only succinctly comment on the ability of the multi-trait model to accommodate the covariance structure of cQTL residuals. The reported results are typical of the model performances in different settings.

Tables 2–5 give the true cQTL effects $\eta^l_{j,k}$ and their estimated values as posterior means $\bar{\eta}^l_{j,k}$ under the two specifications (MD1 and MD2) of the HB multi-trait cQTL model for different missing value scenarios and different values of the prior inclusion probability s^l. In all tables, a bold font is used to indicate the positions where the true or estimated effect was non-zero. The shaded cells indicate false positives or false negatives.

Table 5 | True and estimated (posterior means) cQTL effects under the MD2 version of the HB multi-trait cQTL model with 10% markers and 30% expressions coded as missing and different values of Bernoulli parameter s^l.

Location (j, k) of the effect $\eta_{j,k}$		Size of the regulatory effects $\eta^l_{j,k}$, $l = \{M, E, ME\}$								
Pair j	Trait k	η^M_j			η^E_j			η^{ME}_j		
		True value	Estimated		True	Estimated		True value	Estimated	
			$s^l = 0.013$	$s^l = 0.09$		$s^l = 0.013$	$s^l = 0.09$		$s^l = 0.013$	$s^l = 0.09$
4	1	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	**6**	0	**0**
	3	0	0	0	0	0	0	0	0	0
9	1	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	**0.31**	0	0	0
	3	0	0	0	0	0	0	0	0	0
14	1	0	0	0	**6**	**5.49**	**5.42**	0	0	0
	2	0	0	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0	0
18	1	0	0	**0.49**	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0	0
	3	0	0	0	**6**	**5.60**	**5.48**	0	0	0
19	1	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0	**−0.52**
	3	0	0	0	0	0	0	0	0	0
22	1	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	**1.88**	0
	3	0	0	0	0	0	0	0	0	0
24	1	**6**	**0**	**1.21**	0	0	**−0.47**	0	0	0
	2	**6**	**0**	**0.20**	0	0	**−1.02**	0	0	**7.61**
	3	**6**	**0.2**	**1.09**	0	0	**−0.55**	0	0	0

A bold font is used to indicate positions where the true or estimated effect was non-zero ($|\tilde{\eta}^l_{j,k}| \geqslant 0.2$). The shaded cells indicate false positives and false negatives.

The results shown in **Tables 2–5** suggest that the multi-trait cQTL model is more effective at identifying cQTLs. The better performance was observed when s^l was set to be small (0.013), owing presumably to the fact that a lower s^l value implies a stronger constraint on the presence of effect, which may prevent redundant effects from showing up. Conversely, a higher s^l value increases the model propensity to false discovery, but may result in more accurate estimates since in this case the effect sizes experience relatively less shrinkage. However, the point of cQTL analysis is variable selection rather than estimation, meaning that the accurate estimation of the effects is not essential. It is also clear from our results that the mixed phenotype × expression η^{ME} effects are the most difficult to identify. Moreover, the FDR proved to increase with the proportion of missing data. Finally, cQTL identification was more effective under the MD1 specification. The model was particularly ineffective at identifying the mixed regression parameter η^{ME} under the MD2 specification, and was more prone to false discovery than under the MD1 specification. A potential explanation for this propensity to false discovery is the lack of constraint in the missing data model $E_{i,j} \sim N(0, \sigma_0^2)$ under MD2.

More interestingly, the cQTL residual covariance matrix **S** was estimated to an appreciable degree of accuracy under the multi-trait model in case of correlated cQTL residuals. The estimates (posterior medians) of the components of **S** under the MD1 specification of the multi-trait model with 10% of markers and 10% expression coded as missing and $s^l = 0.013$ was $\tilde{S} \approx \begin{pmatrix} 122.6 & & \\ 96.9 & 83 & \\ 100.9 & 80 & 87.5 \end{pmatrix}$, which is fairly close to the simulated values.

In single-trait cQTL analyses, the results were comparable across the three traits. **Tables 6–9** give the results for single-trait analysis of trait 1 based on simulated data under uncorrelated cQTL residuals. These results are representative of the full set of (uncorrelated data) results from the single-trait cQTL analysis.

In single-trait cQTL analysis too, the model performed better under the MD1 specification when s^l was low (0.0033 and 0.013). Although the model was capable of detecting roughly all true effects under both the MD1 and MD2 specifications, the FDR was relatively higher under the MD2 specification. The number of false positives appeared to increase with the proportion of missing expressions under both specifications.

Table 6 | True and estimated (posterior means) cQTL effects in the analysis of Trait 1 using the MD1 version of the HB single-trait cQTL model with (A) 10% markers and 10% expressions coded as missing and (B) 10% markers and 30% expressions coded as missing, and different values of Bernoulli parameter s^I.

Pair j	Size of the regulatory effects $\eta^I_{j,1}$, $I = \{M, E, ME\}$								
	η^M_j			η^E_j			η^{ME}_j		
		Estimated			Estimated			Estimated	
	True value	$s^I = 0.0033$	$s^I = 0.013$	True value	$s^I = 0.0033$	$s^I = 0.013$	True value	$s^I = 0.0033$	$s^I = 0.013$
(A) MD1, TRAIT 1, 10% G, 10% E CODED AS MISSING									
3	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0
14	0	0	0	**6**	**0**	**0.69**	0	0	0
17	0	0	0	0	0	0	0	0	0
24	**6**	**0.65**	**0.67**	0	**−0.46**	**−0.38**	0	0	0
(B) MD1, TRAIT 1, 10% G, 30% E CODED AS MISSING									
14	0	0	0	**6**	**0.78**	**2.6**	0	0	0
17	0	0	0	0	0	0	0	0	0
24	**6**	**0.56**	**0.72**	0	0	**−0.55**	0	0	0

A bold font is used to indicate positions where the true or estimated effect was non-zero. The shaded cells indicate false positives and false negatives.

Table 7 | True and estimated (posterior means) cQTL effects in the analysis of Trait 1 using the MD2 version of the HB single-trait cQTL model with (A) 10% markers and 10% expressions coded as missing and (B) 10% markers and 10% expressions coded as missing, and different values of s^I.

Pair j	Size of the regulatory effects $\eta^I_{j,1}$, $I = \{M, E, ME\}$								
	η^M_j			η^E_j			η^{ME}_j		
	True value	Estimated		True	Estimated		True value	Estimated	
		$s^I = 0.0033$	$s^I = 0.013$		$s^I = 0.0033$	$s^I = 0.013$		$s^I = 0.0033$	$s^I = 0.013$
(A) MD2, TRAIT 1, 10% G, 10% E CODED AS MISSING									
3	0	**0.21**	**0.30**	0	0	0	0	0	0
6	0	**0.28**	**0.41**	0	0	0	0	0	0
9	0	0	**0.26**	0	0	0	0	0	0
14	0	0	0	**6**	**0**	**0.46**	0	0	0
17	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0		0
24	**6**	**1.55**	**2.40**	0	0	0	0	0	0
(B) MD2, TRAIT 1, 10% G, 30% E CODED AS MISSING									
3	0	0	**0.23**	0	0	0	0	0	0
6	0	**0.24**	**0.20**	0	0	0	0	0	0
9	0	0	**0.39**	0	0	0	0	0	0
14	0	0	0	**6**	**0.63**	**2.58**	0	0	0
17	0	0	0	0	0	0	0	0	0
24	**6**	**1.38**	**2.38**	0	−0.21	−0.23	0	0	0

A bold font is used to indicate positions where the true or estimated effect was non-zero. The shaded cells indicate false positives and false negatives.

DISCUSSION

The conceptual description of the new HB multi-trait cQTL model was presented in this paper and it provides a promising framework for integrating molecular markers and gene transcript levels to dissect the genetic architecture of complex clinical traits. Our results demonstrate that the multi-trait approach enhances the power and should be considered seriously in cQTL mapping framework. Because of its conceptual nature, it is worth

Table 8 | True and estimated (posterior means) cQTL effects in the analysis of Trait 1 using the MD1 version of the HB single-trait cQTL model with (A) 10% markers and 10% expressions coded as missing and (B) 10% markers and 30% expressions coded as missing, and for $s^l = 0.09$.

Pair j	Size of the regulatory effects $\eta^l_{j,1}$, $l = \{M, E, ME\}$					
	η^M_j		η^E_j		η^{ME}_j	
	True value	**Estimated**	**True value**	**Estimated**	**True value**	**Estimated**
(A) MD1 TRAIT 1, WITH 10%G, 10%E CODED AS MISSING						
1	0	0	0	0	0	0
3	0	**0.23**	0	0	0	0
6	0	0	0	0	0	0
12	0	0	0	0	0	0
14	0	0	**6**	**2.36**	0	0
17	0	**0.70**	0	0	0	0
18	0	**0.22**	0	0	0	0
19	0	0	0	0	0	0
22	0	0	0	0	0	**0.23**
23	0	0	0	0	0	**0.27**
24	**6**	**0.83**	0	**−0.45**	0	0
(B) MD1 TRAIT 1, WITH 10%G, 30%E CODED AS MISSING						
3	0	**0.21**	0	0	0	0
6	0	0	0	0	0	0
12	0	0	0	0	0	0
13	0	0	0	0	0	0
14	0	0	**6**	**4.49**	0	0
16	0	0	0	0	0	0
17	0	**0.63**	0	**−0.26**	0	0
18	0	**0.22**	0	0	0	0
19	0	0	0	0	0	0
22	0	0	0	0	0	0
23	0	0	0	0	0	**0.20**
24	**6**	**0.71**	0	**−0.39**	0	0

emphasizing that practical and scalable implementations of the method are beyond the scope of this paper. Often considerable amount of missing marker and expression data may occur at random positions in the data matrix with higher missing rate for expressions than for marker genotypes. The analyses here were based on two missing data scenarios with either 10% of both marker genotypes $G_{i,j}$ and gene expressions $E_{i,j}$ coded as missing, or 10% of marker genotypes $G_{i,j}$ and 30% of gene expressions $E_{i,j}$ coded as missing. For both the multi-trait and single-trait approaches, we considered two different model specifications depending on the way the missing expression data were modeled namely, MD1 (shown in **Figure 1**) where $E_{i,j} \sim N(I_j \mu_j A_j G_{i,j}, \sigma_0^2)$, and MD2 where $E_{i,j} \sim N(0, \sigma_0^2)$.

Under both MD1 and MD2 specifications, the priors on the inclusion indicators, $I^l_{k,j}|s^l$, for the cQTL effects were defined as $I^l_{k,j}|s^l \sim Bern(s^l)$, and different pre-specified values were used for prior inclusion probability s^l, including $s^l = 0.013 = 1/(3 \times 25)$, which assumes fewer non-zero indicator elements (i.e., a sparser model) than the true simulated value 0.08, and the slightly larger value $s^l = 0.09$.

For the sake of comparison, we also analyzed each trait separately through the single-trait cQTL model with three different values for s^l namely, 0.0033, 0.013, and 0.09. The single-trait analyses were confined to simulated data under uncorrelated cQTL residuals.

The multi-trait model performed better under the MD1 specification in terms of identifying cQTLs for small to moderate (10–30%) proportion of missing expression data, and tended to produce fewer false positives. Under the MD2 specification, the model failed consistently to identify the mixed genotype × expression effects η^{ME} when s^l was large (0.09), regardless of the amount of missing data, whilst under the same conditions. Overall, the MD1 version of the model was capable of identifying 75% of the mixed genotype × expression effects.

The out-performance of the MD1 version of the model over its MD2 counterpart in terms of the power to identify the non-zero cQTL effects and an overall lower rate of false positives was also observed in single-trait cQTL analyses. This suggests that the handling of missing values can make a difference to the model performance.

The multi-trait HB cQTL model showed over its single-trait counterpart an increased power of identifying cQTLs with a lower rate of false positives. The covariance structure of cQTL residuals

Table 9 | True and estimated (posterior means) cQTL effects in the analysis of Trait 1 using the MD2 version of the HB single-trait cQTL model with (A) 10% markers and 10% expressions coded as missing and (B) 10% markers and 30% missing expressions coded as missing, and for Bernoulli parameter $s^I = 0.09$.

Pair j	Size of the regulatory effects $\eta^I_{j,1}$, $I = \{M, E, ME\}$					
	η^M_j		η^E_j		η^{ME}_j	
	True value	Estimated	True value	Estimated	True value	Estimated
(A) MD2, TRAIT 1, 10% G, 10% E CODED AS MISSING						
1	0	0	0	0	0	0
3	0	**0.53**	0	0	0	0
6	0	**0.89**	0	0	0	0
9	0	**1.08**	0	0	0	**−0.46**
12	0	0	0	0	0	0
14	0	0	**6**	**1.57**	0	0
15	0	0	0	0	0	0
16	0	0	0	0	0	0
17	0	**0.66**	0	0	0	0
18	0	**0.39**	0	0	0	0
21	0	0	0	0	0	0
22	0	**0.22**	0	0	0	0
24	**6**	**2.40**	0	0	0	0
(B) MD2, TRAIT 1, 10% G, 30% E CODED AS MISSING						
1	0	0	0	0	0	0
3	0	**0.50**	0	0	0	0
6	0	**0.33**	0	0	0	0
9	0	**2.03**	0	0	0	0
12	0	0	0	0	0	**0.23**
13	0	0	0	0	0	0
14	0	0	**6**	**4.64**	0	0
15	0	0	0	0	0	0
16	0	**−0.21**	0	0	0	0
17	0	**0.47**	0	**−0.32**	0	0
18	0	**0.28**	0	0	0	0
19	0	0	0	0	0	0
22	0	**0.24**	0	0	0	0
24	**6**	**2.72**	0	0	0	0

was also estimated under the multi-trait model to a fair degree of accuracy. The multi-trait approach to cQTL analysis is valuable for addressing a number of practical challenges arising in the presence of correlated phenotypic traits, as is the case for many complex disease syndromes like asthma (e.g., Kim and Xing, 2009). Moreover, a multi-trait model provides a framework for investigating a number of biologically interesting hypotheses involving multiple traits, such as pleiotropy. In conclusion, the HB approach to multi-trait cQTL analysis holds great promises for elucidating the underlying biology of complex clinical traits.

ACKNOWLEDGMENTS

We are grateful to two anonymous referees for their constructive comments. This work was supported by a research grant from the Academy of Finland and the University of Helsinki's research funds.

REFERENCES

Borevitz, J. O., Liang, D., Plouffe, D., Chang, H. S., Zhu, T., Weigel, D., Berry, C. C., Winzeler, E., and Chory, J. (2003). Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Res.* 13, 513–523.

Breitling, R., Li, Y., Tesson, B. M., Fu, J., Wu, C., Wiltshire, T., Gerrits, A., Bystrykh, L. V., de Haan, G., Su, A. I., and Jansen, R. C. (2008). Genetical genomics: spotlight on QTL hotspots. *PLoS Genet.* 4, e1000232. doi:10.1371/journal.pgen.1000232

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

Cheung, V. G., Spielman, R. S., Ewens, K., Weber, T. M., Morley, M., and Burdick, J. T. (2005). Mapping determinants of human gene expression by regional and genome-wide association. *Nature* 437, 1365–1369.

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

Drake, T. A., Schadt, E. E., and Lusis, A. J. (2006). Integrating genetic and gene expression data: application to cardiovascular and metabolic trait in mice. *Mamm. Genome* 17, 466–479.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*, 2nd Edn. New York: Chapman and Hall.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (eds). (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

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

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

Jiang, C., and Zeng, Z.-B. (1995). Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* 140, 1111–1127.

Kendziorski, C. M., Chen, M., Yuan, M., Lan, H., and Attie, A. D. (2006). Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics* 62, 19–27.

Kim, S., and Xing, E. P. (2009). Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS Genet.* 5, e1000587. doi:10.1371/journal.pgen.1000587

Lee, E., Cho, S., Kim, K., and Park, T. (2009). An integrated approach to infer causal associations among gene expression, genotype variation, and disease. *Genomics* 94, 269–277.

Li, H., and Deng, H. (2010). Systems genetics, bioinformatics and eQTL mapping. *Genetica* 138, 915–924.

Liu, B., de la Fuente, A., and Hoeschele, I. (2008). Gene network inference via structural equation modeling in genetical genomics experiments. *Genetics* 178, 1763–1776.

Liu, J., Liu, Y., Liu, X., and Deng, H.-W. (2007). Bayesian mapping of QTL for multiple complex traits with the use of variance components. *Am. J. Hum. Genet.* 81, 304–320.

Mackay, T. F. C. (2009). Q&A: genetic analysis of quantitative traits. *J. Biol.* 8, 23.

Mutshinda, C. M., O'Hara, R. B., and Woiwod, I. P. (2008). Species abundance dynamics under neutral assumptions: a Bayesian approach to the controversy. *Funct. Ecol.* 22, 340–347.

Mutshinda, C. M., and Sillanpää, M. J. (2010). Extended Bayesian LASSO for multiple quantitative trait loci mapping and unobserved phenotype prediction. *Genetics* 186, 1067–1075.

Mutshinda, C. M., and Sillanpää, M. J. (2011). Bayesian shrinkage analysis of QTLs under shape-adaptive shrinkage priors, and accurate re-estimation of genetic effects. *Heredity* 107, 405–412.

Pikkuhookana, P., and Sillanpää, M. J. (2009). Correcting for relatedness in Bayesian models for genomic data association analysis. *Heredity* 103, 223–237.

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

Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., Guhathakurta, D., Sieberts, S. K., Monks, S. A., Reitman, M., Zhang, C., Lum, P. Y., Leonardson, A., Thieringer, R., Metzger, J. M., Yang, L., Castle, J., Zhu, H., Kash, S. F., Drake, T. A., Sachs, A., and Lusis, A. J. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.* 37, 710–717.

Schadt, E. E., Monks, S. A., and Friend, S. H. (2003). A new paradigm for drug discovery: integrating clinical, genetic, genomic and molecular phenotype data to identify drug targets. *Biochem. Soc. Trans.* 31, 437–443.

Sillanpää, M. J., and Noykova, N. (2008). Hierarchical modeling of clinical and expression quantitative trait loci. *Heredity* 101, 271–284.

Thomas, A., O'Hara, R. B., Ligges, U., and Sturtz, S. (2006). Making BUGS open. *R News* 6, 17–21.

Wittkopp, P. J. (2005). Genomic sources of regulatory variation in cis and in trans. *Cell. Mol. Life Sci.* 62, 1779–1783.

Wu, C., Delano, D. L., Mitro, N., Su, S. V., Janes, J., McClurg, P., Batalov, S., Welch, G. L., Zhang, J., Orth, A. P., Walker, J. R., Glynne, R. J., Cooke, M. P., Takahashi, J. S., Shimomura, K., Kohsaka, A., Bass, J., Saez, E., Wiltshire, T., and Su, A. I. (2008). Gene set enrichment in eQTL data identifies novel annotations and pathway regulators. *PLoS Genet.* 4, e1000070. doi:10.1371/journal.pgen.1000070

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

Zou, W., Aylor, D. L., and Zeng, Z.-B. (2007). eQTL Viewer: visualizing how sequence variation affects genome-wide transcription. *BMC Bioinformatics* 8, 7. doi:10.1186/1471-2105-8-7

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 December 2011; accepted: 12 May 2012; published online: 06 June 2012.

Citation: Mutshinda CM, Noykova N and Sillanpää MJ (2012) A hierarchical Bayesian approach to multi-trait clinical quantitative trait locus modeling. Front. Gene. ***3****:97. doi: 10.3389/fgene.2012.00097*

This article was submitted to Frontiers in Statistical Genetics and Methodology, a specialty of Frontiers in Genetics.