Heredity (2009) 103, 223–237

www.nature.com/hdy

ORIGINAL ARTICLE

Correcting for relatedness in Bayesian models for genomic data association analysis

P Pikkuhookana and MJ Sillanpää
Department of Mathematics and Statistics, Rolf Nevanlinna Institute, University of Helsinki, Helsinki, Finland

For small pedigrees, the issue of correcting for known or estimated relatedness structure in population-based Bayesian multilocus association analysis is considered. Two such relatedness corrections: [1] a random term arising from the infinite polygenic model and [2] a fixed covariate following the *class D* model of Bonney, are compared with the case of no correction using both simulated and real marker and gene-expression data from lymphoblastoid cell lines from four CEPH families. This comparison is performed with clinical quantitative trait locus (cQTL) models—multilocus association models where marker data and expression levels of gene transcripts as well as possible genotype × expression interaction terms are jointly used to explain quantitative trait variation. We found out that regardless of having a correction term in the model, the cQTL-models fit a few extra small-effect components (similar to finite polygenic models) which itself serves as a relatedness correction. For small data and small heritability one may use the covariate model, which clearly outperforms the infinite polygenic model in small data examples.
Heredity (2009) **103,** 223–237; doi:10.1038/hdy.2009.56; published online 20 May 2009

Keywords: Bayes; cQTL; multilocus association analysis; SNP; gene expression; family structure

Introduction

Population-based marker–phenotype association studies suffer from confounding due to population structure and cryptic relatedness (residual dependencies) that have not been observed or accounted for among the study subjects (Lander and Schork 1994; Yu *et al*., 2006; Iwata *et al*., 2007). The same applies to expression–phenotype association studies (Gibson, 2003; Kraft and Horvath, 2003) and clinical quantitative trait locus (cQTL) studies where genotypes and gene expressions are simultaneously used to study the association with the phenotype (Hoti and Sillanpää, 2006; Bhattacharjee and Sillanpää, 2009; Sillanpää and Noykova, 2008). The significance of this problem in human association studies is currently a subject of considerable debate (Marchini *et al*., 2004; Devlin *et al*., 2004; Hinds *et al*., 2004; Helgason *et al*., 2005; Clayton *et al*., 2005; Voight and Pritchard, 2005; Setakis *et al*., 2006; Zhao *et al*., 2007). If no pedigree/ancestry information is available, there are different approaches to estimate the unobserved structure of population or of the pedigree using neutral molecular markers (Pritchard *et al*., 2000; Blouin, 2003; Excoffier and Heckel, 2006; Weir *et al*., 2006; Gasbarra *et al*., 2007; Bink *et al*., 2008). In many cases, however, exact information specifying the interrelations between individuals may be available. This is so, for example when data has been ascertained specifically from families or from pedigrees (see Visscher *et al*., 2008).

Correspondence: Dr MJ Sillanpää, Department of Mathematics and Statistics, Rolf Nevanlinna Institute, University of Helsinki, P.O. Box 68, FIN-00014 Helsinki, Finland.
E-mail: mjs@rolf.helsinki.fi
Received 8 September 2008; revised 9 April 2009; accepted 14 April 2009; published online 20 May 2009

Robust methods have been developed especially for family-based association studies (Gauderman *et al*., 1999; Zhao, 2000; Knapp and Becker, 2003; Chen and Abecasis, 2007) and case–control association testing with related individuals (Thornton and McPeek, 2007). To correct for population structure in population-based association analyses, one can for example adjust the *P*-value (Devlin and Roeder, 1999), include the population term in the association model (Yu *et al*., 2006; Zhao *et al*., 2007) or use a principal component approach (Price *et al*., 2006). Similarly, for known or estimated relatedness, there are many ways to include such information in the association model. One approach to correct for cryptic relatedness is to include relationships in the form of a (covariance) matrix into the association model in the studies of marker–phenotype association (see George and Elston, 1987; Kennedy *et al*., 1992; Jannink *et al*., 2001; Yu *et al*., 2006) or in the studies of expression–phenotype association (see Lu *et al*., 2004). One can incorporate an additive relationship matrix (the covariance structure of a multivariate normal distribution) either to residuals or use a specific random term (arising from the infinite polygenic model) in the regression model. In linkage studies the same term appears in the role of the genetic background. Such covariance structure takes care of the dependencies between the study subjects. Another approach approximates such structure by having the phenotype of the parents, sibs and the spouse of the subject as covariates in the regression model and assumes independence for residuals (Bonney, 1986). This autoregressive structure does not model the true underlying dependence structure, but has been shown to perform well and to account for confounding with single locus models. These two (polygenic and covariate) correction terms have been studied and used earlier

only in a single-gene association model and the testing framework. Thus their properties for Bayesian multilocus association models (for example, Kilpikari and Sillanpää, 2003; Sillanpää and Bhattacharjee, 2005) are largely unknown.

Modelling phenotype with both gene expression data and marker data could be advantageous and provides more information because marker data is stable in comparison with time- and tissue-dependent gene expression data (O'Hara, 2006; West *et al*., 2006). Although this view has been confirmed in simulations (Hoti and Sillanpää, 2006; Sillanpää and Noykova, 2008) it remains arguable with real data (Bhattacharjee *et al*., 2008; Bhattacharjee and Sillanpää 2009). We compare here how these two relatedness corrections work together with Bayesian model-based multilocus association using both marker and gene expression data. To do this, we have modified the cQTL-model of Hoti and Sillanpää (2006) for SNPs and pedigree data. Our emphasis is on a collection of small pedigrees, as cryptic relatedness has a negligible effect in large outbred populations, especially when the sample size increases (Voight and Pritchard, 2005). We consider only a small amount ($\sim$5%) of missing data here (cf. Sillanpää and Noykova, 2008).

Model

cQTL model

Let N_M be the number of SNP (single nucleotide polymorphism) markers and N_E be the number of gene expression transcripts. Our data consists of continuous phenotypes $y=(y_1,\ldots,y_n)^t$, SNP marker genotypes $m=(m_{1,1},m_{1,2},\ldots,m_{n,N_M})$ and gene expression measurements $x=(x_{1,1},x_{1,2},\ldots,x_{n,N_E})^t$ from n individuals in the known pedigrees collected from a single population. We assume that each individual has its own observation (array) on gene expression made at single time point. For cases with multiple populations, see the Discussion section. Here, we let y_i denote the observed continuous phenotype of the ith individual and summarize the genotypes as $z_{i,j,1}=1_{\{m_{i,j}=AA\}}$, $z_{i,j,2}=1_{\{m_{i,j}=AB\}}$ and $z_{i,j,3}=1_{\{m_{i,j}=BB\}}$, where $z_{i,j,k}$ denotes the indicator of kth genotype at the SNP j for individual i. The gene expression measurement j for the individual i will be denoted by $x_{i,j}$. We want to emphasize that we assume that gene expression levels are available (with some missing entries) for each study subject. We closely follow the notation in Hoti and Sillanpää (2006) and assume that gene expression measurements are normalized (Quackenbush, 2001; Butte, 2002) and transformed suitably beforehand, so that sample distribution of the majority of the genes is approximately standard normal. Moreover, we assume that N_E and N_M are relatively small (a few hundreds at most). To form candidates for genotype $\times$ expression interactions, we assume that some markers are *a priori* associated (that is, possibly have a regulatory effect) with some gene expression measurements. This prior information on the pairing of markers and expressions may be obtained from previous, independent studies, or could be based on known pathways or proximity of their genomic location. We refer to them as marker–gene pairs (see Hoti and Sillanpää, 2006). We allow multiple expressions to be associated with a single marker, but not the other way around. Let $\tilde{x}_{i,j}=x_{i,g_j}$ and $\tilde{z}_{i,j,k}=z_{i,s_j,k}$ be the expression measurement and genotype indicator for some pair (g_j,s_j) so that for individual i $\tilde{x}_{i,j}$ is the gene expression measurement of gene g_j and $\tilde{z}_{i,j,k}$ is the indicator of genotype k at SNP s_j. The number of these previously assigned pairs is N_{ME}. We consider the following linear model for a continuous phenotype

$$\begin{aligned} y_i =& \mu+\sum_{j=1}^{N_M}\sum_{k=1}^{3} I_j^M\alpha_{j,k}z_{i,j,k}+\sum_{j=1}^{N_E} I_j^E\beta_j x_{i,j} \\ &+\sum_{j=1}^{N_{ME}}\sum_{k=1}^{3} I_j^{ME}\gamma_{j,k}\tilde{z}_{i,j,k}\tilde{x}_{i,j}+F_i+\varepsilon_i, \end{aligned} \tag{1}$$

where μ is the population mean and $\varepsilon_i\sim N(0,\ \sigma_0^2)$ is a normally distributed residual term with mean zero and variance σ_0^2. F_i denotes a correction term, which takes into account the family structure and the dependence between family members. The linear regression coefficient (effect) of genotype k at marker j is $\alpha_{j,k}$, coefficient of expression effect is β_j and coefficient of the interaction effect is $\gamma_{j,k}$. Unlike in Hoti and Sillanpää (2006), each genetic component, marker, expression or interaction, has its own indicator variable, I_j^M, I_j^E or I_j^{ME}, respectively. For our motivation for the use of indicators, see the Discussion section. For indicators, the value one corresponds to the inclusion and value zero to the exclusion of the genetic component in the model. Obviously SNP markers exhibit three genotypes and we use an overparameterized model, so for each marker and for each marker–gene pair (that is, marker $\times$ expression interaction), there is a single indicator variable and three effect coefficients. (We allow the first coefficients ($\alpha_{j,1}$, $\gamma_{j,1}$) at each locus j to be unconstrained in our model unlike that in Hoti and Sillanpää (2006)). We can identify differences (genotypic contrasts) as functions of posteriors afterwards from the Markov chain Monte Carlo (MCMC) sample or from the MCMC point estimates. As in Hoti and Sillanpää (2006) we can write the genetic data of individual i as the vector

$$X_i=(z_{i,1,1},\ldots,z_{i,N_M,3},x_{i,1},\ldots,x_{i,N_E},\tilde{z}_{i,1,1}\tilde{x}_{i,1},\ldots,\tilde{z}_{i,N_{ME},3}\tilde{x}_{i,N_{ME}}),$$

and the vector containing $N=3N_M+N_E+3N_{ME}$ unknown effects is denoted by

$$\begin{aligned}\theta &= (\theta_1,\theta_2,\ldots,\theta_N)\\ &= (\alpha_{1,1},\ldots,\alpha_{N_M,3},\beta_1,\ldots,\beta_{N_E},\gamma_{1,1},\ldots,\gamma_{N_{ME},3}).\end{aligned}$$

In addition there are $N_M+N_E+N_{ME}$ indicator variables. To create the vector that contains the indicator variables for all genetic effects, we need to arrange the indicators into the vector containing N elements

$$\begin{aligned}I^* =& (I_1^*,I_2^*,\ldots,I_N^*)=(I_1^M,I_1^M,I_1^M,\ldots,I_{N_M}^M,I_{N_M}^M,I_{N_M}^M,I_1^E,\ldots,\\ &I_{N_E}^E,I_1^{ME},I_1^{ME},I_1^{ME},\ldots,I_{N_{ME}}^{ME},I_{N_{ME}}^{ME},I_{N_{ME}}^{ME})\end{aligned}$$

Now we can rewrite the linear cQTL-model (1) as

$$y_i=\mu+\sum_j^N I_j^*\theta_j X_{i,j}+F_i+\varepsilon_i.$$

Infinite polygenic model

One approach to taking family structure into account in the model is to add a random individual effect, whose

correlation structure would follow the degree of relationship between individuals. In the cQTL-model (1), the term F_i represents the additive effects of the polygenes on individual i, which arise from the combined action of infinitely many loci whose individual contributions cannot be distinguished (Yi and Xu, 2000; Jannink *et al.*, 2001). The additive polygenic effects $F=(F_1,\ldots,F_n)^t$ are distributed as multivariate normal with known covariance structure, $F \sim MVN(\tilde{0}, 2\Phi\sigma_F^2)$. Here, $\tilde{0}$ is a $n\times 1$ vector of zeros and Φ is a $n\times n$ matrix of kinship coefficients among individuals based on pedigree information and σ_F^2 is the additive variance of the polygenes (George and Elston, 1987; Kennedy *et al.*, 1992; Jannink *et al.*, 2001; Monks *et al.*, 2004). The kinship coefficient between two individuals is the expected probability that homologous genes taken randomly from their genomes are identical by descent from common ancestors in the given pedigree (Lynch and Walsh, 1998). In the breeding literature, F_i's are called breeding values and 2Φ the additive relationship matrix (Henderson, 1976). The structure of the matrix Φ is block-diagonal once the individuals are arranged by the families and no remote shared ancestry among the families is assumed. For simplicity, we assume no inbreeding and omit the dominance component.

Regression covariates

Another approach to describing family dependencies is to add the phenotypes of the relatives as covariates (fixed effects) into the model. Then for individual i we can write F_i in the cQTL-model (1) as

$$F_i = \rho_f y_{f_i}{}' + \rho_m y_{m_i}{}' + \rho_s y_{s_i}{}' + \rho_{os}\sum_{j\in os_i} y_j',$$

where y'denotes deviations of the phenotypes from their empirical mean, subscript f_i refers to father, m_i denotes mother, s_i denotes spouse if she/he appears earlier in the data set and os_i is the set of sibs of individual i appearing earlier in the data set (Bonney, 1986; Thomas, 2004). Here, ρ_f, ρ_m, ρ_s and ρ_{os} are respective regression coefficients, which can be written in the vector form as $\rho=(\rho_f, \rho_m, \rho_s, \rho_{os})$. Bonney (1986) referred to this as the *class D* model. This kind of model structure can be seen as an approximation of polygenic background (Thomas, 2004).

Hierarchical model

Prior distributions

We need to specify prior distributions for the unknown parameters. We allow each genetic effect in vector $(\theta=\theta_1, \theta_2, \ldots, \theta_N)$ to have its own variance parameter (Xu, 2003; Hoti and Sillanpää, 2006). For the genetic effects θ, we assign prior $p(\theta|\sigma^2)=\prod_{j=1}^N p(\theta_j|\sigma_j^2)$, where the functional form of $p(\theta_j\,|\,\sigma_j^2)$ is a normal density with the mean zero and the effect-specific variance σ_j^2. We assigned to σ_j^2 the Jeffreys' prior $p(\sigma_j^2)\propto 1/\sigma_j^2$, which together with effect-specific variances induce sparseness into the model (Xu, 2003; Hoti and Sillanpää, 2006). By sparseness, we mean that most of the effects are zero or almost zero. For details of implementation, see the Estimation section in Appendix A. There is also another source of sparseness in our model—indicator variables. For indicator variables I, we assign the Bernoulli distribution with parameter $s=P(I_j=1)\ll 0.5$, which is the prior selection probability for a candidate to be included in the model (that is, $I=1$). For parameter s we give values $1/N_M$, $1/N_E$ or $1/N_{ME}$, for markers, expressions and their interactions, respectively. This is equivalent to assuming *a priori* that there is one selected effect for each type of genetic component. We treat priors $p(\theta)$ and $p(I)$ independently (Kuo and Mallick, 1998; Sillanpää and Bhattacharjee, 2005; Sillanpää and Noykova, 2008). The prior for μ is $p(\mu)\propto 1$, and prior density for $\sigma_0^2=\mathrm{var}(\varepsilon_i)$ is $p(\sigma_0^2)\propto 1/\sigma_0^2$. As a prior for polygenic effects, we use the multivariate normal density. That is

$$p(F|\sigma_F^2) = \frac{1}{(2\pi)^{n/2}|2\Phi\sigma_F^2|^{1/2}}\exp\left[-\tfrac{1}{2}F'(2\Phi\sigma_F^2)^{-1}F\right],$$

where $F=(F_1, \ldots, F_n)$ is a vector of polygenic effects, Φ is a matrix of kinship coefficients among individuals, σ_F^2 is additive polygenic variance with prior $p(\sigma_F^2)\propto 1/\sigma_F^2$ and $|2\Phi\sigma_F^2|$ is the determinant of the covariance matrix. Now, under the additive polygenic model, the joint prior is $p(\theta, I, F, \mu, \sigma^2) = p(I)p(\theta|\sigma^2)p(F|\sigma_F^2)p(\mu)p(\sigma^2)$, where $p(\sigma^2)=p(\sigma_F^2)\prod_{j=0}^N p(\sigma_j^2)$. For the regression covariate model we replace $p(F\,|\,\sigma_F^2)$ with $p(\rho)=\prod_{j\in\{f,m,s,os\}} p(\rho_j)$, where $p(\rho_j)$ is a normal density function with the mean zero and variance 1000, and $p(\sigma^2)=\prod_{j=0}^N p(\sigma_j^2)$.

Missing data model

We assume data are missing at random (Rubin, 1976) and treat missing values as unknown random variables in Bayesian inference. Thus, we need to specify a prior distribution for missing observations. Denote the complete genetic data with no missing values by $D=\{m, x\}$. The observed genetic data with possibly some missing values is denoted by $D^-=\{m^-, x^-\}$. Recall that the gene expression measurements were assumed to be normalized beforehand. Prior distribution for missing gene expression measurement is assumed simply to be a standard normal distribution (cf. for a major gene model, see Sillanpää and Noykova, 2008). Even if in this model the polygenic basis is assumed for gene expressions, we omit (genetic) dependencies from parents. In the prior distribution for missing genotypes we take into account the genotypic values of individuals' parents, but omit the recombination aspect because we do not utilize linkage information in our association model here (see the Discussion section). The joint probability distribution of the marker j over individuals is given by

$$p(m_j) = \prod_{i\in Founders} p(m_{i,j}) \prod_{\substack{i\in non-\\ Founders}} p(m_{i,j}|m_{m,j}, m_{f,j}),$$

where $m_j=(m_{1,j},\ldots,m_{n,j})^t$ is the genotype pattern at marker j. The first product is over the prior probabilities of the genotypes of founders, and the second is over transmission probabilities of genotypes of non-founders and $m_{m,j}$ and $m_{f,j}$ are the genotypes of mother and father of individual i, respectively. Transmission probabilities $p(m_{i,j}\,|\,m_{m,j},m_{f,j})$ follow the Mendelian rules of inheritance. Note that although it seems that there are dependencies in transmission only downwards the pedigree, in practise there are also upward dependencies due to total probability. The genotypes of the founders are thought of as being drawn from the population with uniform allele frequencies. Then, the prior density function of the

genetic data is

$$p(D) \propto \prod_{j=1}^{N_M} \left(\prod_{i \in Founders} p(m_{i,j}) \prod_{\substack{i \in non- \\ Founders}} p(m_{i,j}|m_{m,j}, m_{f,j}) \right) \times \prod_{j=1}^{N_E} \left(\prod_{i=1}^{n} p(x_{i,j}) \right).$$

For details of implementation, see the Estimation section in Appendix A.

Posterior distributions

In Bayesian analysis, marginal posterior distributions for the parameters are derived from the prior distributions and the likelihood of the data. Using Bayes formula, the joint posterior density of the model parameters conditional on phenotypic and genetic data is given by

$$p(\theta, I, F, \mu, \sigma^2, D|y, D^-) \propto p(\theta, I, F, \mu, \sigma^2) \times p(D)p(D^-|D)p(y|\theta, I, F, \mu, \sigma^2, D)$$

where $p(\theta, I, F, \mu, \sigma^2)$ is the density function of the joint prior distribution of parameters $\{\theta, I, F, \mu, \sigma^2\}$, $p(D)$ is the prior density function of the complete genetic data, $p(D^- \mid D)$ is the mass probability function of the observed genetic data D^- conditional on the complete genetic data D (that is, is the indicator function and takes value 1 only when D^- is consistent with D and is zero otherwise) and $p(y|\theta, I, F, \mu, \sigma^2, D) = \prod_{i=1}^{n} p(y_i|\theta, I, F, \mu, \sigma^2, D)$ is the likelihood of the phenotype data, where

$$p(y_i|\theta, I, F, \mu, \sigma^2, D) \propto \frac{1}{\sqrt{\sigma_0^2}} \times \exp\left(-\frac{1}{2\sigma_0^2} \left(y_i - \mu - \sum_{j=1}^{N} I_j^* \theta_j X_{i,j} - F_i \right)^2 \right).$$

Examples of cQTL analysis with family data

To compare corrections for family data with cQTL-model (1), we analyse a few data sets with three-generation pedigrees in the presence of missing data. First, we analyse two simulated data sets with known genetic effects and then a real CEPH family data that have been used in previous studies (Kraft *et al*., 2003; Schadt *et al*., 2003). We also consider average performance (assessed by analysing 25 data replicates). The simulated data is an example of a large data set (210 individuals) with loosely correlated genetic components and real data is an example of a small sample size (58 individuals) with highly correlated genetic components. We first compare how the two correction terms (infinite polygenic model and covariate model) perform against no correction term (model for unrelated individuals) with family data, which has either single or multiple simulated trait-influencing components and compare two correction terms with the real CEPH data. Finally, in simulated data replicates, we consider only marker–phenotype association and compare three methods using 25 marker data sets with three trait-loci.

Simulations

We simulated family data consisting of molecular markers, gene expression level measurements and a continuous phenotype. Our simulation procedure follows the procedure of Hoti and Sillanpää (2006), where expression levels are first generated conditionally on markers, and phenotypes are then generated conditionally on them both. The main difference is that we use real SNP marker data on families as a starting point. We want to emphasize that this approach is able to generate realistic dependence structures for markers as well as expressions. Real marker data was obtained from the CEPH genotype database (Dausset *et al*., 1990). Fifteen families from the CEPH/Utah family collection were selected with the family identifiers 1334, 1340, 1345, 1346, 1349, 1350, 1358, 1362, 1375, 1377, 1408, 1418, 1421, 1424 and 1477. Selection criteria were large number of children and large number of genotypes available for all three generations (cf. Monks *et al*., 2004). In total, the families represent 210 individuals. We selected 52 SNPs from eight different chromosomes, based on the availability of genotypes (not too many missing values) and the property that selected markers was not highly dependent (closely linked) to one another (Table 1). We also required that the markers are in Hardy–Weinberg equilibrium and that minor allele frequency (MAF) was not less than 5%.

Simulating SNP genotypes: First, we needed to complete missing genotypes in CEPH families, as our simulation procedure for expression and phenotype data (below) necessitates complete SNPs. There were less than 5% of genotypes missing among the set of selected markers. Genotypes were missing entirely on two individuals and

Table 1 ID-numbers of SNPs selected from 8 chromosomes. There were a couple of hundreds of SNPs available on each chromosome in the database

Chr. 1	*Chr. 2*	*Chr. 3*	*Chr. 4*	*Chr. 5*	*Chr. 6*	*Chr. 7*	*Chr. 8*
TSC0000177	TSC0036676	TSC0234875	TSC0029679	TSC0029334	TSC0003041	TSC0100414	TSC0049269
TSC0078081	TSC0249367	TSC0303065	TSC0211587	TSC0109491	TSC0111315	TSC0230024	TSC0144304
TSC0167604	TSC0536527	TSC0787063	TSC0329568	TSC0627529	TSC0379931	TSC0355119	TSC0336892
TSC0289448	TSC0652130	TSC1051029	TSC0454238	TSC0645321	TSC0457233	TSC0846687	TSC0706315
TSC0393286	TSC0679932	TSC1588387	TSC0674462	TSC0669621	TSC0931610	TSC1057539	
TSC0458980	TSC0896550		TSC1079646	TSC0746604	TSC1794266	TSC1630115	
TSC0980921	TSC1202072		TSC1548043	TSC0876755			
TSC1200125	TSC1766295			TSC1777520			

The SNPs were selected so that none of the selected SNPs was included in the list of close markers of another selected SNP in the CEPH genotypic database.

some SNPs were not available for a couple of families. We sampled the missing genotypes of the founders from the population of equal allele frequencies conditionally on the progeny. This allowed us to consistently fill data (missing genotypes) downwards through the pedigree. Genotypes were drawn according to the Mendelian transmission probabilities. In this process, we omitted recombination probabilities, but took fully into account that every missing genotype depends on genotypes of the parents on the same SNP marker.

Simulating expression levels: Conditionally on an individual's genotype on each marker (m_j), we simulated three gene expression measurements ($x_{3\times j-2}$, $x_{3\times j-1}$, $x_{3\times j}$). The first two of these ($x_{3\times j-2}$, $x_{3\times j-1}$) had constant probability to have *in cis* effect and the third gene ($x_{3\times j}$) was set to have no regulatory effects with probability one. *A priori* (before simulating actual values) we divided markers into three *in cis* effect groups. One-third of the markers (m_j) were assigned an *in cis* effect on gene expressions if the genotype was homozygote (*AA*), one-third had an *in cis* effect if genotype was heterozygote (*AB*) and final third had an *in cis* effect if the genotype was homozygote (*BB*). The decision that the marker actually exhibits the pre-specified *in cis* effect was made with probability 0.3, which is in line with previous estimates (Jansen and Nap, 2004; Morley *et al*., 2004). Gene expression measurements ($x_{3\times j}$) at positions with no *in cis* effect, and gene expression measurements ($x_{3\times j-2}$ and $x_{3\times j-1}$), in absence of *in cis* effect, were simulated from the distribution $N(0,1)$. In presence of *in cis* effect, the expression value of the one (*in cis*) gene ($x_{3\times j-2}$) assigned on current marker j was simulated from the distribution $N(2,1)$ and expression value of another gene ($x_{3\times j-1}$) assigned on the same marker was simulated from the distribution $N(-2,1)$ (see Figure 1).

Simulating phenotypes: Excluding simultaneously active components at each marker–gene pair, genetic components can be divided into six subtypes, depending on their effect on phenotype and whether an *in cis* effect is present or absent in marker–gene pairs. Following Hoti and Sillanpää (2006), we denote these subtypes as genotype effect without *in cis* effect (*G*), genotype effect with *in cis* effect (*iG*), gene expression effect without *in cis* effect (*E*), gene expression effect with *in cis* effect (*iE*), genotype × gene expression effect without *in cis* effect (*GE*) and genotype × gene expression effect with *in cis* effect (*iGE*). Continuous phenotypes are constructed as a linear combination of six underlying genetic components and the polygene.

$$y_i = \sum_{k=1}^{3} a_{1,k} z_{i,S_1,k} + \sum_{k=1}^{3} a_{2,k} z_{i,S_2,k} + a_3 x_{i,S_3} + a_4 x_{i,S_4} + \sum_{k=1}^{3} a_{5,k} x_{i,S_5} z_{i,S_5,k} + \sum_{k=1}^{3} a_{6,k} x_{i,S_6} z_{i,S_6,k} + g_i + \varepsilon_i,$$

where $s_1,\ldots, s_6$ are indexes of influential marker–gene pairs of types *G*, *iG*, *E*, *iE*, *GE* and *iGE*, respectively, $z_{i,j,k}$ is

marker m_j	expression $x_{3\times j-2}$	P(*in cis* effect)=0.3	$N(2,1)$
			$N(0,1)$
(in group 1, where *in cis* effect) is attached to genotype AA)	expression $x_{3\times j-1}$	P(*in cis* effect)=0.3	$N(-2,1)$
			$N(0,1)$
	expression $x_{3\times j}$	P(no *in cis* effect)=1.0	$N(0,1)$

marker m_j	expression $x_{3\times j-2}$	P(*in cis* effect)=0.3	$N(2,1)$
			$N(0,1)$
(in group 2, where *in cis* effect) is attached to genotype AB)	expression $x_{3\times j-1}$	P(*in cis* effect)=0.3	$N(-2,1)$
			$N(0,1)$
	expression $x_{3\times j}$	P(no *in cis* effect)=1.0	$N(0,1)$

marker m_j	expression $x_{3\times j-2}$	P(*in cis* effect)=0.3	$N(2,1)$
			$N(0,1)$
(in group 3, where *in cis* effect) is attached to genotype BB)	expression $x_{3\times j-1}$	P(*in cis* effect)=0.3	$N(-2,1)$
			$N(0,1)$
	expression $x_{3\times j}$	P(no *in cis* effect)=1.0	$N(0,1)$

Figure 1 Distributions from which the gene expression measurements for different genotypic groups were generated. Probability of *in cis* effect is 0.3.

Table 2 Simulated genetic components and their effect sizes

Marker/gene	*Simulated effects*
$S_1 = 9$	$a_{1,1} = -2$, $a_{1,2} = 2$, $a_{1,3} = 6$
$S_2 = 52$	$a_{2,1} = -4$, $a_{2,2} = 1$, $a_{2,3} = 4$
$S_3 = 86$	$a_3 = 2$
$S_4 = 167$	$a_4 = 5$
$S_5 = 274$	$a_{5,1} = -2$, $a_{5,2} = 1$, $a_{5,3} = 5$
$S_6 = 215$	$a_{6,1} = 3$, $a_{6,2} = -1$, $a_{6,3} = -5$

Approximately 9% of phenotypic variance was thought to result from the polygenic component and approximately 60% was due to simulated genetic components, so heritability of the phenotype was ≈ 0.72.

the indicator of genotype k at the marker j for the individual i, $x_{i,l}$ is the gene expression value of gene l for the individual i, and the environmental residual ε_i is assumed to be normally distributed with mean 0 and variance 1. Polygenic terms g_i are simulated jointly from the multivariate normal distribution with zero mean vector and covariance–variance structure $2\Phi\sigma^2$, where 2Φ describes the additive relationships between CEPH family members and σ^2 is the additive polygenic variance. Here, σ^2 is fixed to some value for the desired degree of heritability due to polygenes. See Table 2 for details of the simulation including values of the coefficients ($a_{1,k}$, $a_{2,k}$, a_3, a_4, $a_{5,k}$, $a_{6,k}$) and variability due to given effects and the polygene. To test how well our method behaves in case of missing data, we randomly discarded 5% of the data on phenotypes, genotypes and expressions.

We also simulated phenotypic data with one underlying genetic component and the polygene. The starting point for the phenotypic simulation was that the genotypic data was identical with the earlier data set and expression levels were simulated similarly. The only influential genetic component was SNP marker $s_1 = 36$ without *in cis* effect. Simulated effects were $a_{1,1} = -2$ and $a_{1,3} = 6$ for the homozygotes and $a_{1,2} = 2$ for the heterozygote. The simulated polygenic component explained approximately 7% and the simulated SNP effects approximately 24% of the phenotypic variance. We also randomly discarded 5% of the phenotypes, genotypes and expression measurements in this data set.

Simulating replicated data sets: To evaluate the average performance of the methods, we simulated 25 phenotypic data sets with the same marker effects and the polygene in each. The genotypic family data was the same as in earlier simulations (210 individuals and 52 SNP markers) and was kept unchanged in all simulations. We simulated three trait loci (SNPs 7, 29 and 36) with their own effect sizes. For SNP $s_1 = 7$, we simulated effects $a_{1,1} = 1$ and $a_{1,3} = 9$ for the homozygotes and $a_{1,2} = 5$ for the heterozygote, for SNP $s_2 = 29$ we simulated effects $a_{2,1} = -3$, $a_{2,3} = 1$ and $a_{2,2} = -1$ and for SNP $s_1 = 36$ we simulated effects $a_{3,1} = -2$, $a_{3,3} = 4$ and $a_{3,2} = 1$. The simulated polygenic component was approximately 17% of the phenotypic variance but varied due to sampling variation in different realizations. Simulated overall heritability varied equally from 0.34 to 0.52 in replicates. Replicated analyses were done for the complete data sets with no missing values.

Real data

We analysed gene expression data of the lymphoblastoid cell lines of 58 individuals from four CEPH families (CEPH/Utah pedigrees 1362, 1375, 1377 and 1408). The original article about the data set is Schadt *et al*. (2003). The sibship data from the same families has been used earlier in Kraft *et al*. (2003) as test data to examine the performance of the FEXAT statistic, which represents a sort of correlation coefficient for family data. Technical details about measuring gene expression in this data set can be found in Schadt *et al*. (2003). CEPH lymphoblastoid cell lines had been cultured and maintained in the log phase of cell growth at least 2 days before harvest (Schadt *et al*., 2003). At the time of measuring the expression, it would be expected that the WNT pathway would be active, because the WNT pathway has been shown to regulate B lymphocyte proliferation (Reya *et al*., 2000). Following Kraft *et al*. (2003), we chose the expression of β-catenin (CTNNB1NM_001904) as a clinical quantitative trait, and expect that in the presence of WNT, levels of the β-catenin (trait) will be associated with factors that can lead to the formation and stabilization of the β-catenin/TCF complex. On the other hand, in the absence of WNT, β-catenin levels will be associated with genes making up the β-catenin destruction complex (Seidensticker and Behrens, 2000).

Gene expression measurements were obtained from the NCBI GenBank. Locations of genes are based on reference assembly. For every gene, we additionally searched the closest available SNP, which is genotyped for these same four CEPH families, using the same criteria as in the simulation analysis. Genotypes were obtained from the CEPH genotype database. Maximum distance between a gene and the closest SNP was 2 361 528 bp, whereas there was no minimum distance, because one SNP was found inside the gene region (Table 3). We omitted individuals who did not have expression measurements at all.

Results

Simulated data

Analysis details and effect summaries: For data sets with 5% of the data missing, we run our models with WinBUGS 1.4.1 using four separate MCMC chains each of length 10 000. For each chain, burn-in was 1000 and thinning 10 (that is, only every 10th MCMC sample was stored), and samples from all chains were combined in MCMC estimation of the parameters. For checking the convergence of each chain, we visually inspected MCMC paths of several parameters. We summarize our results as posterior genetic occupancy probabilities for genetic component j, $P(\textit{occupancy at location } j \mid \textit{data})$, obtained as the proportion of MCMC rounds where the indicator variable I_j is 1, indicating that genetic component j is included in the model. Note that there are as many indicator variables as genetic components ($N_M + N_E + N_{ME}$) with continuous indexing. We also calculated conditional probabilities $Q_{j,k} = P(I_J = 1 \mid I_k = 1, \textit{data})$ for all pairs (j,k) of indicator variables, which showed elevated posterior probabilities (cf. Hoti and Sillanpää, 2006). $Q_{j,k}$ is the posterior probability that the genetic component j is included in the model on the condition that the genetic component k is included in the

Table 3 List of putative genes and their closest available SNP markers

Genes expected to be associated with the β-catenin destruction complex				*Genes expected to be associated with β-catenin/TCF complex*			
Gene	*GenBank accession number for gene*	*Probe name in CEPH database for SNP*	*Distance between gene and SNP*	*Gene*	*GenBank accession number for gene*	*Probe name in CEPH database for SNP*	*Distance between gene and SNP*
GSK3B	NM_002093	TSC1051029	2 245 085	LEF1	NM_016269	TSC0281761	1 838 287
AXIN1	AF009674	TSC0201072	—	TCF4	NM_003199	TSC0540674	992 877
AXIN2	NM_004655	TSC0143579	851 543	CTBP2	NM_001329	TSC0203917	1 154 067
APC	NM_000038	TSC0379412	462 122	WNT11	NM_004626	TSC1012488	1 872 550
DVL1	NM_004421	TSC0903069	2 361 528	WISP2	NM_003881	TSC0417608	118 009
DVL2	NM_004422	TSC0457749	629 621	MAP3K1	AF042838	TSC0362155	226 479
DVL3	NM_004423	TSC1079572	494 783	MAP3K2	NM_006609	TSC0652130	1 888 200
				MAP3K5	NM_005923	TSC0803320	221 290
				MAP3K7	NM_003188	TSC0926016	1 195 963
				MAP3K13	NM_004721	TSC0493500	624 274
				MAP3K8	NM_005204	TSC1055614	1 295 333
				MAP3K6	NM_004672	TSC0110133	1 497 052
				MAP3K14	NM_003954	TSC0031889	2 702 808
				MAP3K4	NM_005922	TSC0919533	738 460
				MAP3K12	NM_006301	TSC0796032	199 773
				MAP3K3	NM_002401	TSC0143579	976 195
				MAP3K11	NM_002419	TSC0919351	2 823 474

Table includes accession numbers of genes and probe names of SNPs. Genes on the left (right) are associated with β-catenin degradation (stabilization), respectively.

model. In our preliminary analysis (results not shown), we have found that sometimes the expression effect may be captured by the interaction term where the same expression measurement is involved. Same kind of complementary behaviour of the effects was present also in Hoti and Sillanpää (2006) and in Sillanpää and Noykova (2008). So, we calculated Q summaries also for expression components, which corresponds to elevated interaction terms and vice versa. The number of genetic components in the model was summarized (in each MCMC round) as the number of indicator variables, which were simultaneously 1. Heritability was estimated by using the formula

$$h^2 \approx \frac{1}{r}\sum_{t=1}^{r}\frac{\sigma_y^{2(t)}-\sigma_e^{2(t)}}{\sigma_y^{2(t)}},$$

where $\sigma_y^{2(t)}$ is phenotypic variance and $\sigma_y^{2(t)}$ $\sigma_e^{2(t)}$ is residual variance at round t, and r is total number of MCMC rounds after burn-in. Note that the estimated phenotypic variance depends on imputed values.

Analysis results: Throughout the paper, we used the threshold 0.1 to determine significant components. For the data with six simulated effects, the model with infinite polygenic correction term found five and the covariate model found six genetic components with elevated posterior occupancy probabilities (Figure 2). One of these ($j=31$), which was found with both models, was actually a false positive, which cannot be explained, with any simulated effects. The expression effects were partly captured by interaction terms so that the expression effect was rarely in the model at the same time as the corresponding interaction term. This can be seen from the Q summaries (Table 4) where the conditional probabilities were less than 0.07 and 0.04 (for an infinite polygenic model and a covariate model, respectively) for all such cases. The same can be seen also from the MCMC paths of the indicator times the effect (Figure 3). Here, $I_j^* \times \theta_j$ (product of the indicator and effect size) shows, that the expression effect ($j=167$) and the genotype × expression interaction ($j=323$) are clearly complementary, indicating that only one of them contributes to the model at a time. Even though the occupancy probabilities for the simulated components of the type E and GE were both smaller than 0.1, they are clearly higher than the occupancy probabilities of the other similar type of components (Figure 2). As illustrated in the Table 5, the highest posterior probability was obtained for the correct number of genetic components in both the infinite polygenic model $P(n_c=6\,|\,data)\approx 0.22$ and in the covariate model $P(n_c=6\,|\,data)\approx 0.23$. However, posterior support was obtained for a wide range of values varying from 2 to 11 and 3 to 11 components in the infinite polygenic model and in the covariate model, respectively. When 5% of the data was artificially coded as missing, run time was approximately 15 h (4 chains) for every 1000 rounds of iterations for both models.

The same data was also analysed with the model which did not include any correction term for the pedigree structure (that is, $F_i=0$ for all i). Surprisingly, the same six effects with elevated occupancy probabilities were found here as in the covariate model analysis. The false positive ($j=31$) showed slightly higher probability in this analysis than with the other two models (Table 6). Again the highest posterior probability was obtained for the correct number of simulated genetic components ($P(n_c=6\,|\,data)\approx 0.24$). The mean posterior estimated number of genetic components included simultaneously in the model was slightly higher for this model than for the other two models (Table 7).

In the infinite polygenic model analysis of data with one simulated effect, the true simulated component was always captured correctly in the model. However, the number of genetic components in the model was clearly overestimated, even though there was no strong evidence for any false positives. There were only two genetic components with occupancy probabilities larger than 0.1 and one of them was false positive, although

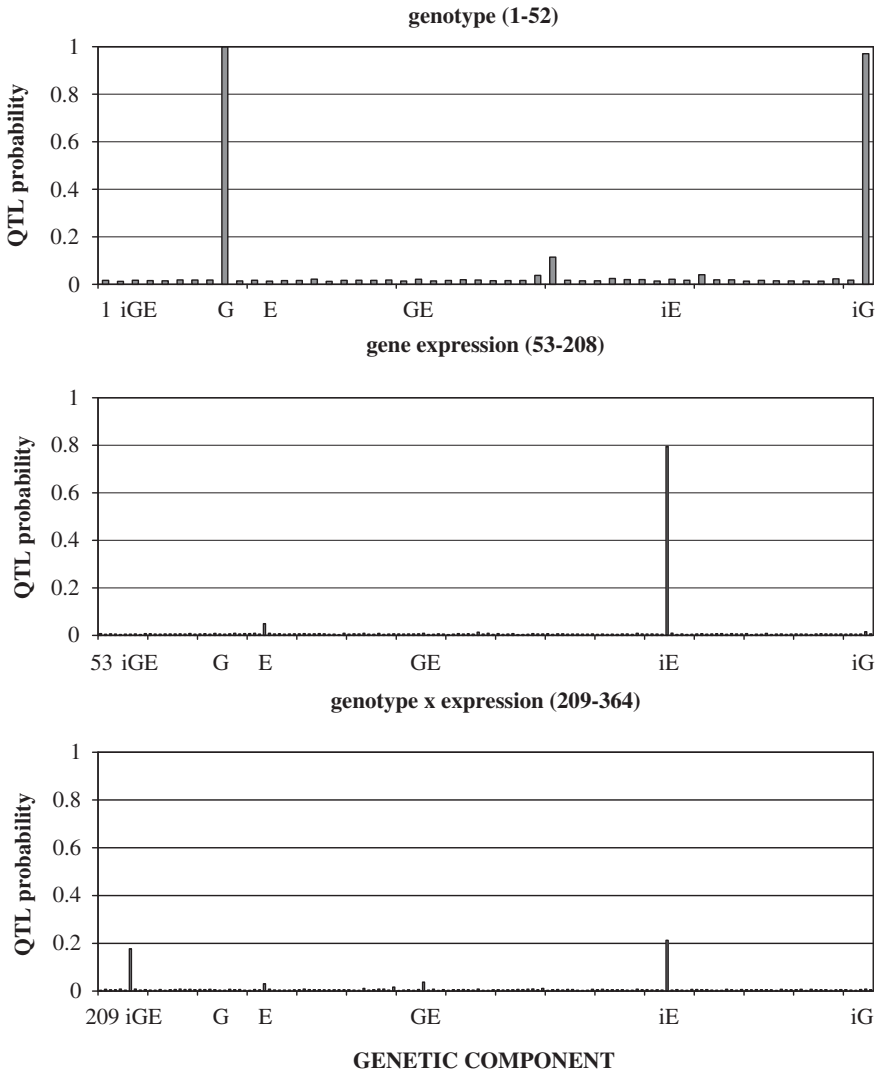

Figure 2 A summary of analysis with the covariate model. The panels contain the estimated posterior occupancy probability for the genotype effects (top), gene expression effects (middle) and genotype × expression effects (bottom). The positions of simulated effects are indicated by a shortcut notation of the effect subtype. The corresponding genetic components are vertically levelled.

Table 4 Pairwise conditional summaries

J	$P(I_j=1 \mid \text{data})$	k										
		9	*31*	*52*	*59*	*86*	*118*	*167*	*215*	*242*	*274*	*323*
9	1.0		0.99	1.0	1.0	1.0	0.97	1.0	1.0	1.0	1.0	1.0
31	0.12	0.11		0.12	0.17	0.12	0.09	0.12	0.06	0.13	0.14	0.09
52	0.97	0.97	0.99		1.0	0.98	0.94	0.97	0.99	0.96	0.99	0.99
59	0.00	0.01	0.01	0.01		0.01	0.00	0.01	**0.00**	0.00	0.01	0.00
86	0.05	0.05	0.05	0.05	0.11		0.06	0.05	0.07	**0.02**	0.02	0.04
118	0.01	0.01	0.01	0.01	0.00	0.01		0.01	0.01	0.01	**0.01**	0.01
167	0.79	0.79	0.83	0.79	0.89	0.83	0.82		0.85	0.88	0.76	**0.04**
215	0.18	0.18	0.10	0.18	0.17	0.26	0.12	0.19		0.16	0.21	0.13
242	0.03	0.03	0.03	0.03	0.00	**0.01**	0.03	0.03	0.03		0.01	0.02
274	0.04	0.04	0.05	0.04	0.06	0.02	0.03	0.04	0.05	0.02		0.04
323	0.21	0.21	0.17	0.22	0.11	0.17	0.18	**0.01**	0.16	0.12	0.24	

The Q-summaries for covariate model. The first column contains indexes j of the genetic components. The second column contains their posterior occupancy probabilities and remaining matrix contains the pairwise conditional probabilities $Q_{j,k}=P(I_J=1 \mid I_k=1, \textit{data})$ where k is given in the top row. Values indicating complementary components are highlighted in bold.

there was some probability mass for as many as nine influential components (Table 5). The covariate model performed slightly better than the infinite polygenic model and was able to include the true simulated component in the model with probability one, whereas there were no other components with occupancy

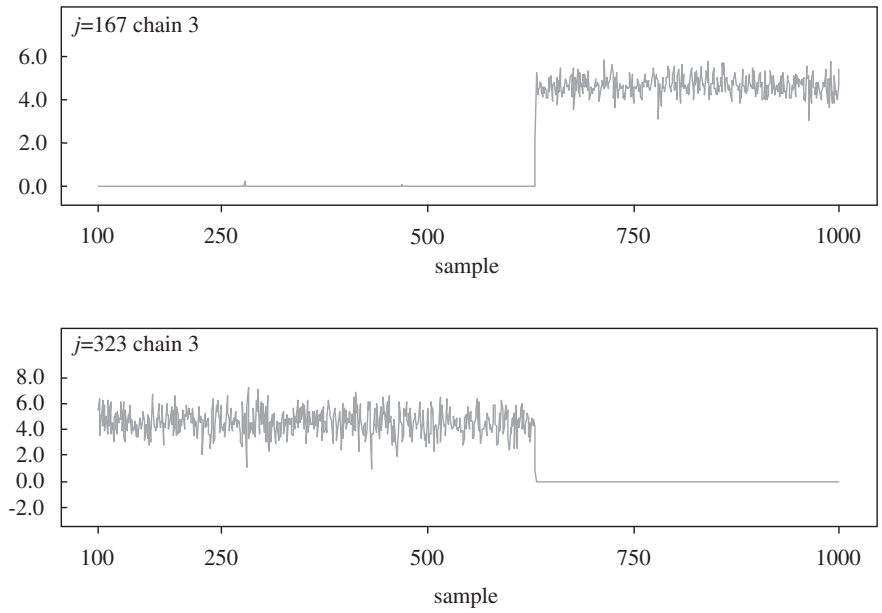

Figure 3 The MCMC paths of the *effect* × *I* for the expression component 167 (top) and the interaction component 323 corresponding to genotype AB (bottom). The burn-in period is 1000 with thinning 10 and has been removed from the MCMC sample before drawing the figure.

Table 5 Number of components

Analysis	n_c													
	0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*
Six simulated effects														
Polygenic model	0.00	0.00	0.01	0.05	0.13	0.21	0.22	0.17	0.11	0.06	0.02	0.01	0.00	0.00
Covariate model	0.00	0.00	0.00	0.04	0.13	0.21	0.23	0.18	0.11	0.05	0.02	0.01	0.00	0.00
No correction	0.00	0.00	0.00	0.03	0.13	0.20	0.24	0.19	0.11	0.06	0.02	0.01	0.00	0.00
One simulated effect														
Polygenic model	0.00	0.05	0.16	0.23	0.23	0.16	0.10	0.05	0.02	0.01	0.00	0.00	0.00	0.00
Covariate model	0.00	0.06	0.15	0.23	0.24	0.16	0.10	0.03	0.01	0.01	0.00	0.00	0.00	0.00
No correction	0.00	0.05	0.16	0.23	0.23	0.17	0.09	0.04	0.02	0.01	0.00	0.00	0.00	0.00

The posterior probabilities of the different number of genetic components in the model (the number of indicator variables being simultaneously 1) estimated from the two simulated data sets.

Table 6 Effect estimates and occupancy probabilities

Location	*Type of effect*	*Genotype*	*Simulated value*	*Polygenic model*		*Covariate model*		*No correction*	
				Effect × *I* *estimate*	$P(I_j=1 \mid data)$	Effect × *I* *estimate*	$P(I_j=1 \mid data)$	Effect × *I* *estimate*	$P(I_j=1 \mid data)$
9	*G*	*AB*	4	0.12 [−0.55, 1.81]	0.97	0.17 [−0.45, 2.20]	1.00	0.15 [−0.54, 2.15]	1.00
		BB	8	7.02 [0, 10.13]		7.51 [4.49, 10.58]		7.32 [4.23, 10.31]	
31		*AB*	0	−0.92 [−9.71, 0]	0.14	−0.75 [−9.10, 0]	0.11	−1.26 [−9.82, 0]	0.18
		BB	0	−0.92 [−9.71, 0]		−0.75 [−9.13, 0]		−1.27 [19.84, 0]	
52	*iG*	*AB*	5	5.41 [0,9.53]	0.82	6.88 [0, 10.08]	0.97	6.47 [0, 9.62]	0.96
		BB	8	5.71 [0, 10.12]		7.19 [0, 10.55]		6.76 [0, 10.04]	
86	*E*		2	0.03 [0, 0]	0.03	0.08 [0, 1.77]	0.05	0.04 [0, 0]	0.03
167	*iE*		5	4,41 [0, 5.54]	0.94	3.69 [0, 5.46]	0.79	3.52 [0, 5.43]	0.75
215	*iGE*	*AB*	−4	−0.85 [−4.84, 0]	0.32	−0.38 [−3.85, 0]	0.18	−0.69 [−4.58, 0]	0.30
		BB	−8	−1.68 [−7.53, 0]		−0.81 [−6.69, 0]		−1.48 [−7.34, 0]	
242		*AB*	0	0.03 [0, 0]	0.02	−0.06 [0, 0.08]	0.03	0.03 [0, 0]	0.02
		BB	0	0.01 [0, 0]		−0.01 [0, 0]		0.01 [0, 0]	
274	*GE*	*AB*	3	0.01 [0, 0]	0.04	0.01 [0, 0]	0.04	0.01 [0, 0]	0.04
		BB	6	0.13 [0, 2.69]		0.14 [0, 3.39]		0.14 [0, 3.01]	
323		*AB*	0	0.19 [0, 4.10]	0.07	0.62 [−0.09, 5.44]	0.21	0.70 [−0.34, 5.25]	0.25
		BB	0	0.20 [0, 4.39]		0.63 [0, 4.94]		0.75 [−0.06, 5.01]	

Simulated and estimated effects (posterior mean and 95% credible interval) under three competing models when constraining the effect of the genotype *AA* to zero. Here, '*Effect* × *I*' refers to $P(effect \times I_j \mid data)$.

Table 7 Point estimates of the number of components

No. of effects	*Mean*	*s.d.*	*2.5%*	*Median*	*97.5%*
Polygenic model	6.03	1.83	3	6	10
Covariate model	6.09	1.71	3	6	10
No correction	6.20	1.72	3	6	10

Posterior estimated number of influential components for the data with six simulated components using the three models.

probabilities larger than 0.1. The model without correction term found the true simulated genotypic component and one genotypic component with occupancy probability 0.1 and one interaction component with occupancy probability 0.098.

Heritability and effect estimates: For the data with six simulated effects the infinite polygenic model underestimated the heritability in its posterior point-estimate whereas the similar estimate from the covariate model was even smaller (Table 8). The 95% Credibility Interval for infinite polygenic model included the true simulated heritability but the CI was wider than for the covariate model or for the model with no correction. The estimates of the effect sizes were also underestimated. It turned out that there was a clear dependence, as expected: the higher the posterior occupancy probability was for the genetic component the more accurate the estimate for the effect was. This was true especially for the expression effects. Table 6 presents the comparison between simulated and estimated genetic effects. For effects, we show only the model-averaged estimate $I \times \theta$, because it is more robust (Ball, 2001) and I appears always together with θ in the model (cf. Sillanpää and Bhattacharjee, 2005). In the table the effect of the genotype AA is constrained to zero to make values more comparable. Posterior estimate for the additive polygenic variance was much smaller than the true simulated value and it seemed in the trace plot nearly zero for most of the MCMC iterations. It is likely that other genetic components (SNPs, expressions and their interactions) captured some of the polygenic variance by subdividing a small amount of variance to be explained by each component. Thus correction term estimates were relatively modest (Table 9). Also, it is likely that the heritability estimate suffered from the fact that some simulated components stayed unselected for most of the MCMC iterations.

Analyses with all three models for the data with one simulated effect also underestimated heritability (Table 8), but the 95% CI included the true simulated value in all models. The estimated polygenic variance behaved the same way as in the data with six simulated effects. The effect estimates were slightly better with the covariate model though the infinite polygenic model and the model without correction also gave good estimates. In general, these estimates were more accurate here than for the data with six simulated components (results not shown).

Simulated data replicates

Analysis results: Replicated analysis of marker data sets gave quite similar results with all three methods. All methods found the same trait loci in almost every data set, but their degree of evidence (the magnitude of signals) was slightly different. The infinite polygenic model underestimated polygenic variance and for some data sets it had difficulties of finding a single mode (converging value) and thus had identifiability problems. As a whole, the infinite polygenic model estimated heritability better than the other two models, but still it underestimated the true heritability almost every time. The model without the correction term found simulated effects more frequently (that is, had better power) than the other two models and the infinite polygenic model had the lowest false-positive rate and false-discovery rate but FPR and FDR were quite similar with all three models (Table 10). The performance of the covariate model was not superior with respect to any summary statistic but performance was still comparable to the other models. As earlier, the higher the posterior occupancy probability was for the genetic component, the more accurate the estimate for the effect was.

Table 8 Heritability estimates and the 95% credible intervals around the posterior mean for two simulated data set with three competing models

Simulation analysis	*Mean*	*95% CI*
Six effects $h^2 \approx 0.72$		
Polygenic model	0.552	[0.387, 0.809]
Covariate model	0.501	[0.369, 0.618]
No correction	0.500	[0.364, 0.617]
One effect $h^2 \approx 0.31$		
Polygenic model	0.247	[0.071, 0.410]
Covariate model	0.237	[0.061, 0.381]
No correction	0.238	[0.061, 0.384]

The true heritability of data with six simulated genetic effects was approximately 0.72 and heritability of one simulated effect was approximately 0.31.

Table 9 Correction term estimates

	Mean	*95% CI*
Polygenic variance	2.078	[0.2663, 10.02]
Covariates		
Father	−0.006	[−0.132, 0.125]
Mother	−0.040	[−0.181, 0.105]
Spouse	0.295	[0.057, 0.529]
Sibs	−0.001	[−0.062, 0.057]

Posterior estimates (mean and 95% credible interval) of polygenic variance and covariate coefficients (father, mother, spouse and sibs) from data with six simulated effects.

Real data

Analysis details: When analysing the real data from the CEPH families, we ran four MCMC chains each of length 50 000 and we allowed only the closest marker to have an interaction with corresponding gene expression in the model. In the prior, we restricted all variance components to be less than our empirically estimated phenotypic variance ($\hat{\sigma}^2 \approx 0.007$). The MCMC sampler under the infinite polygenic model showed poor mixing, which resulted in unreliable (non-converged) estimates.

Table 10 Averaged effect estimates and occupancy probabilities of replicated data analysis: Simulated and estimated effects (posterior means) of trait loci under three competing models when constraining the effect of the genotype *AA* to zero

Location	*Genotype*	*Simulated value*	*Polygenic model*		*Covariate model*		*No correction*	
			Effect × I *estimate*	$P(I_j = 1 \mid data)$	Effect × I *estimate*	$P(I_j = 1 \mid data)$	Effect × I *estimate*	$P(I_j = 1 \mid data)$
7	*AB*	4	3.29	0.86	2.97	0.81	3.14	0.90
	BB	8	6.73		6.03		6.71	
29	*AB*	2	0.14	0.07	0.18	0.07	0.24	0.09
	BB	4	0.25		0.25		0.37	
36	*AB*	3	0.91	0.46	0.93	0.43	1.18	0.54
	BB	6	2.31		2.12		2.74	
Average	FNR (%)		34.7		33.3		30.7	
Average	FPR (%)		0.4		0.6		0.5	
Average	FDR (%)		6.7		9.0		8.0	
Average	Heritability	0.43	0.29	[0.07, 0.50]	0.19	[0.01, 0.34]	0.19	[0.01, 0.34]

Here '*Effect* × *I*' refers to $P(effect \times I_j \mid data)$. False-negative rate $FNR = n_{fn}/(n_{tp}+n_{fn})$, false-positive rate $FPR = n_{fp}/(n_{fp}+n_{tn})$, false-discovery rate $FDR = n_{fp}/(n_{fp}+n_{tp})$, and heritability estimate were averaged over analyses of 25 simulated data sets. Here, n_{tp} and n_{fp} are numbers of true and false positives and n_{tn} and n_{fn} are numbers of true and false negatives, respectively.

The MCMC chains of several parameters were stuck in some parts of the parameter space for many iterations and posterior estimates were different and depended on the initial values of the different MCMC runs. In addition, the infinite polygenic model had clear difficulties in separating (identifying) the polygenic variance and the residual variance from each other. During MCMC iterations, most of the time the value of the polygenic variance dominated that of the residual variance which was zero or almost zero, but sometimes this was swapped the other way round. Both these issues probably arise due to the small number of individuals in the data and therefore it is safest not to estimate the variance components from such small data sets (see Misztal, 1996; Burton *et al*., 1999).

Analysis results: The MCMC estimation under the covariate model did not show any problems with mixing, but could not capture any significant genetic effects either. Every genetic effect occurs in the model with almost equal probability, the largest probability (≈ 0.068) was found for the SNP close to gene GSK3B.

In a roundtable discussion (Kass *et al*., 1998) Neal stated that prior constraints may cause convergence problems for Markov chains, so we loosened our prior restriction with genetic variance components in MCMC estimation and allowed them also to have values larger than the phenotypic variance. After this change the covariate model produced slightly elevated posterior probability (≈ 0.130) for the effect of marker × gene expression interaction for gene LEF1. Probabilities for the rest of the effects varied in range (0.019, 0.072). In earlier studies (Behrens *et al*., 1996; Huber *et al*., 1996) LEF1 has been shown to interact with β-catenin, which is an important effector of the WNT-signaling pathway. Together these two proteins mediate a transcriptional response to WNT signalling (Reya *et al*., 2000).

Discussion

In the population-based association analysis of quantitative traits, the use of relatives provides a competitive alternative for a sample of unrelated individuals (Visscher *et al*., 2008). In such cases, the use of a correction term is important in single-gene models to avoid false positives due to the resemblance of individuals (Yu *et al*., 2006; Iwata *et al*., 2007). Two approaches for taking the pedigree structure into account in a model-based multilocus association were presented and compared here with the approach of no correction. In principle, one can easily include a large pedigree in a covariate model. To allow larger pedigrees in the infinite polygenic model, Damgaard (2007) has suggested prior transformation of the kinship matrix to improve the mixing properties of the WinBUGS sampler. However, because we have concentrated on reasonably small pedigrees, we did not apply such a transformation here. Also application of Lin (1999) and Thomas (1992) provide natural samplers for larger pedigrees (see Waldmann, 2009).

Use of indicator variables

Initially, we began by adding a correction term which takes into account the pedigree structure to the model of Hoti and Sillanpää (2006), which does not include any indicator variables. Generally, the model found genetic components quite well, but the heritability estimate had a tendency to become highly inflated (being almost one). We found out that this overestimation was due to the cumulative effect of many negligible genetic effects (at insignificant components) which each contributed very little to the cumulative variance of genetic effects (results not shown). When we added indicator variables into the model (as explained in the Model section), the heritability estimate was affected only by the genetic variances of significant components, whereas the other variances were truly zero (cf. method BayesB in Meuwissen *et al*., 2001). This change in the model structure brought the heritability estimates down from one. It is important to note that Hoti and Sillanpää (2006) obtained good estimates for heritability with their model even without indicators. One reason for different behaviour in Hoti and Sillanpää (2006) and in our

implementation here might be that we made our analysis with WinBUGS, where we had to restrict our flat priors to certain region, which had to be narrow to prevent computational overflows and maintain numerical stability (see Appendix A). This restriction led to the situation where the variance parameters cannot be exactly zero.

Analyses of simulated data

When analysing data with six simulated effects using the infinite polygenic model, our estimate for additive polygenic variance was much smaller than the true simulated value. On the other hand, the estimated number of influential genetic components had some support for being larger than the true number of simulated effects. We found out that these additional effects were all small in size. We suppose that this phenomenon occurs because our model approximates polygenic variance in a similar way as the finite polygenic model (FPM). FPM was first proposed by Thompson and Skolnick (1977) and it describes the genetic (polygenic) covariance among pedigree members by a finite number of unlinked small-effect quantitative trait loci (Du *et al*., 1999; Du and Hoeschele, 2000). Briefly, the correctly identified genetic components and a few extra components together seem to fit (explain) most of the polygenic structure of the data leaving only a small amount of polygenic variance to be explained by the infinite polygenic component. Our model is more flexible than FPM, because FPM assumes a constant number of equal-sized genetic effects when approximating the polygenic structure, whereas our model estimates the number of components and their effects simultaneously from the data. The running of the covariate model with the same data led to the slightly smaller heritability estimate and the same amount of significant genetic effects. Like the infinite polygenic model, also here the multiple markers (and expressions) took the role of polygenic inheritance. Moreover, this performance of the marker effects here is also closely related to the genomic selection (see Meuwissen *et al*., 2001; Calus and Veerkamp, 2007) where the sum of the marker effects is used to model polygenic variation.

In the data with one simulated effect, both the infinite polygenic model and the covariate model favoured more than one influential component in each MCMC iteration. However, these additional components had negligible effects, which were very small in size. This gave further support to the fact that the polygenic inheritance is captured mostly by extra loci in multilocus association models where the effects of multiple loci/components are considered in the model simultaneously.

Analysis of simulated data replicates

These replicated marker data analyses showed us that the infinite polygenic model is very sensitive (in the sense of sometimes providing good estimates and sometimes poor estimates) on particular data in estimating several variance components. All models provided quite similar results, which makes the definition of the best performing method difficult. However, the unpredictable performance of the infinite polygenic model makes it less attractable.

Analysis of real data

Real data analysis with the infinite polygenic model had difficulties in separating polygenic and residual variance during the estimation. This may imply that individual components here also explain/approximate polygenic variance quite well and that the remaining variability cannot be partitioned into two distinct variance components. Also the amount of the data was rather small, so estimating the variance components is not reliable (see Burton *et al*., 1999; Misztal, 1996).

There are several reasons why our approach did not lead to any significant genetic effects on the real data analysis. One is the amount of the data (four families), which was quite small. In addition, it is likely that the heritability of the expression trait is also small. Usually, small amount of data does not matter if (1) heritability is large enough, and vice versa, and if (2) components/candidates are independent. Here, the candidates were especially selected as members from the single WNT pathway, in which case they are evidently highly correlated, which again makes it difficult to do model selection among them using multilocus association models. Our method tries to find a sparse set of trait-associated components at the same time, whereas due to correlatedness, selected components may vary at each MCMC iteration. This may have been the cause of almost equal posterior probabilities for all genetic effects.

In contrast, Kraft *et al*. (2003) used a single-gene test, where the association of a single-gene was tested at a time. The high correlation between candidates in such circumstances could mean that one being significant is the same as them all being significant which may explain the differences between the results. On the other hand, group-based testing would have provided an interesting alternative (Goeman *et al*., 2004).

Model extensions and MCMC estimation

The linkage information of SNPs provided by the pedigree was omitted in our model for the missing data. Linkage can be added into our model so that genotypes of the closely linked loci have dependency structure, by modelling haplotypes and their recombinations as in oligogenic analysis (for example, Heath, 1997; Uimari and Sillanpää, 2001). In principle, the pedigree information also allows an extension to models with combined association and linkage (Fulker *et al*., 1999; George *et al*., 1999; Abecasis *et al*., 2000; Perez-Enciso, 2003). In the case with many missing genotypes in a single family, one must keep in mind that WinBUGS uses the single-site Gibbs sampler in updating missing genotypes. Missing genotypes in consecutive generations can cause the single-site Gibbs sampler to be reducible (Cannings and Sheehan, 2002). In that case one configuration can never be reached, once the other configuration has been assigned earlier. Our model for missing expressions could be extended by utilizing information on *cis*- and *trans*-acting markers (Sillanpää and Noykova, 2008) or by assuming correlated expressions among related individuals to an extent that reflects the heritability. Our model could also be extended to include the correction term for population structure as in Yu *et al*. (2006). However, WinBUGS analysis with multiple populations may be too demanding so that other implementations need to be considered. On the other hand, in light of our results, it is likely that multilocus association models can self-correct population structure similarly as they did for family structure here. Actually, Setakis *et al*. (2006) found this to

be true for population structure in their study by using the binary phenotype and logistic regression, and Iwata *et al*. (2007) for the multilocus association analysis of quantitative traits, see also Iwata *et al*. (2009). In any case, further inspection is needed on the role (importance) of the correction terms in multilocus association models.

There are two sources of sparseness in our model. One results from using Jeffreys' prior and the other from the use of indicator variables (see O'Hara and Sillanpää, 2009). Based on our experiments, Jeffreys' prior dominates the other source of sparseness. It seems that the prior selection probability s has only modest influence on the posterior and that the degree of sparseness here is similar to that which would be obtained from Jeffreys' prior alone. The great benefit of using indicator variables is that they can produce occupancy probabilities directly. Xu (2003) also used Jeffreys' prior to induce sparseness in his model, which did not produce occupancy probabilities for the components. In the model of Hoti and Sillanpää (2006) occupancy probabilities were calculated afterwards for standardized effects using a pre-specified threshold value.

Based on our limited experiments carried out here, Bayesian multilocus models without correction seem to be a flexible tool in association analysis even if there are dependencies among study individuals. When there are many candidate components, they can automatically take residual dependencies into account without producing a large amount of false positives. However, further inspection is needed to clarify when there are enough candidates and data, for it to be safe to leave out the correction term from the cQTL-model. For population structure, Iwata *et al*. (2007) found that use of a correction term (in two-genotype data) systematically seemed to provide some additional advantages over self-correction (the use of a multilocus model without a correction term). It seems that the model without the correction term performs quite similarly with the models, which take into account the pedigree structure. If one, however, wants to use the model with correction we found that if the heritability or the number of individuals is quite small, the use of a covariate model is then preferable. In addition, the covariate model provides a framework to include phenotype information from ungenotyped parents to the analysis (cf. Purcell *et al*., 2005). Nevertheless, the use of the model without the correction term gives satisfactory results when several candidate components are studied in the model simultaneously.

The model specification codes (written in WinBUGS) used in this article are freely available for research purposes from the authors upon request.

Acknowledgements

We are grateful to Bob O'Hara, Andrew Thomas, Petri Koistinen and Crispin Mutshinda Mwanza for discussions and constructive comments on the paper. This work was supported by a research grant (202324) from the Academy of Finland.

Electronic-database information

CEPH genotype database, http://www.cephb.fr/cephdb/

NCBI GenBank, http://www.ncbi.nlm.nih.gov/Genbank/

Gene expression omnibus, http://www.ncbi.nlm.nih.gov/geo/

References

Abecasis GR, Cardon LR, Cookson WOC (2000). A general test of association for quantitative traits in nuclear families. *Am J Hum Genet* **66**: 279–292.

Ball RD (2001). Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian information criterion. *Genetics* **159**: 1351–1364.

Behrens J, von Kries JP, Kühl M, Bruhn L, Wedlich D, Grosschedl R *et al.* (1996). Functional interaction of β-catenin with the transcription factor LEF-1. *Nature* **382**: 638–642.

Bhattacharjee M, Botting CH, Sillanpää MJ (2008). Bayesian biomarker identification based on marker-expression-proteomics data. *Genomics* **92**: 384–392.

Bhattacharjee M, Sillanpää MJ (2009). Bayesian joint disease-marker-expression analysis applied to clinical characteristics of chronic fatigue syndrome. In: McConnell P, Lim S, Cuticchia AJ (eds). *Methods of Microarray Data Analysis VI*. CreateSpace Publishing: Scotts Valley, California. pp 15–34.

Bink MCAM, Anderson AD, van de Weg WE Thompson EA (2008). Comparison of marker-based pairwise relatedness estimators on a pedigreed plant population. *Theor Appl Genet* **117**: 843–855.

Blouin MS (2003). DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends Ecol Evol* **18**: 503–511.

Bonney GE (1986). Regressive logistic models for familial disease and other binary traits. *Biometrics* **42**: 611–625.

ter Braak CJF, Boer MP, Bink MCAM (2005). Extending Xu's Bayesian model for estimating polygenic effects using markers of the entire genome. *Genetics* **170**: 1435–1438.

Burton P, Tiller K, Gurrin L, Cookson W, Musk A, Palmer LJ (1999). Genetic variance components analysis for binary phenotypes using generalized linear mixed models (GLMMs) and Gibbs sampling. *Genet Epidemiol* **17**: 118–140.

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

Calus MPL, Veerkamp RF (2007). Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. *J Anim Breed Genet* **124**: 362–368.

Cannings C, Sheehan NA (2002). On a misconception about irreducibility of the single-site Gibbs sampler in a pedigree application. *Genetics* **162**: 993–996.

Cemgil AT, Févotte S, Godsill CJ (2007). Variational and stochastic inference for Bayesian source separation. *Digital Signal Process* **17**: 891–913.

Chen W-M, Abecasis GR (2007). Family-based association tests for genomewide association scans. *Am J Hum Genet* **81**: 913–926.

Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM *et al*. (2005). Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat Genet* **37**: 1243–1246.

Damgaard LH (2007). Technical note: how to use Winbugs to draw inferences in animal models. *J Anim Sci* **85**: 1363–1368.

Dausset J, Cann H, Cohen D, Lathrop M, Lalouel JM, White R (1990). Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. *Genomics* **6**: 575–577.

Devlin B, Bacanu SA, Roeder K (2004). Genomic control to the extreme. *Nat Genet* **36**: 1129–1130.

Devlin B, Roeder K (1999). Genomic control for association studies. *Biometrics* **55**: 997–1004.

Du F-X, Hoeschele I (2000). Estimation of additive, dominance and epistatic variance components using finite locus models implemented with a single-site Gibbs and a descent graph sampler. *Genet Res* **76**: 187–198.

Du F-X, Hoeschele I, Gage-Lahti KM (1999). Estimation of additive and dominance variance components in finite polygenic models and complex pedigrees. *Genet Res* **74**: 179–187.

Excoffier L, Heckel G (2006). Computer programs for population genetics data analysis: a survival guide. *Nat Rev Genet* **7**: 745–758.

Fulker DW, Cherny SS, Sham PC, Hewitt JK (1999). Combined linkage and association sib-pair analysis for quantitative traits. *Am J Hum Genet* **64**: 259–267.

Gasbarra D, Pirinen M, Sillanpää MJ, Salmela E, Arjas E (2007). Estimating genealogies from unlinked marker data: A Bayesian approach. *Theor Pop Biol* **72**: 305–322.

Gauderman WJ, Witte JS, Thomas DC (1999). Family-based association studies. *J Natl Cancer Inst* **26**: 31–37.

Gelman A, Carlin JBStern HS, Rubin DB (2004). *Bayesian Data Analysis* 2nd edn. Chapman and Hall, London.

George V, Elston RC (1987). Testing the association between polymorphic markers and quantitative traits in pedigrees. *Genet Epidemiol* **4**: 193–201.

George V, Tiwari HT, Zhu X, Elston RC (1999). A test of transmission/disequilibrium for quantitative traits in pedigree data, by multiple regression. *Am J Hum Genet* **65**: 236–245.

Gibson G (2003). Population genomics: celebrating individual expression. *Heredity* **90**: 1–2.

Gilks WR, Thomas A, Spiegelhalter DJ (1994). A language and program for complex Bayesian modelling. *Statistician* **43**: 169–178.

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

Heath SC (1997). Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *Am J Hum Genet* **61**: 748–760.

Helgason A, Yngvadottir B, Hrafnkelsson B, Gulcher J, Stefansson K (2005). An Icelandic example of the impact of population structure on association studies. *Nat Genet* **37**: 90–95.

Henderson CR (1976). A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics* **32**: 69–83.

Hinds DA, Stokowski RP, Patil N, Konvicka K, Kershenobich D, Cox DR *et al.* (2004). Matching strategies for genetic association studies in structured populations. *Am J Hum Genet* **74**: 317–325.

Hopert JP, Casella G (1996). The effect of improper priors on Gibbs sampling in hierarchical mixed models. *J Am Stat Assoc* **91**: 1461–1473.

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

Huber O, Korn R, McLaughlin J, Ohsugi M, Herrmann BG, Kemler R. (1996). Nuclear localization of beta-catenin by interaction with transcription factor LEF-1. *Mech Dev* **59**: 3–10.

Iwata H, Ebana K, Fukuoka S, Jannink J-L, Hayashi T (2009). Bayesian multilocus association mapping on ordinal and censored traits and its application to the analysis of genetic variation among *Oryza sativa* L. germplasms. *Theor Appl Genet* **118**: 865–880.

Iwata H, Uga Y, Yoshioka Y, Ebana K, Hayashi T (2007). Bayesian association mapping of multiple quantitative trait loci and its application to the analysis of genetic variation among *Oryza sativa* L. germplasms. *Theor Appl Genet* **114**: 1437–1449.

Jannink J-L, Bink MCAM, Jansen RC (2001). Using complex plant pedigrees to map valuable genes. *Trends Plant Sci* **6**: 337–342.

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

Kass RE, Carlin BP, Gelman A, Neal RM (1998). Markov Chain Monte Carlo in practice: A roundtable discussion. *Am Stat* **52**: 93–100.

Kennedy BW, Quinton M, van Arendonk JAM (1992). Estimation of effects of single genes on quantitative traits. *J Anim Sci* **70**: 2000–2012.

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

Knapp M, Becker T (2003). Family-based association analysis with tightly linked markers. *Hum Hered* **56**: 2–9.

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

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

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

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

Lin S (1999). Monte Carlo Bayesian methods for quantitative traits. *Comp Stat Data Anal* **31**: 89–108.

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

Lynch M, Walsh B (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Associates: Sunderland, MA.

Marchini J, Cardon LR, Phillips MS, Donnelly P (2004). The effects of human population structure on large genetic association studies. *Nat Genet* **36**: 512–517.

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

Misztal I (1996). Estimation of variance components with large-scale dominance models. *J Dairy Sci* **80**: 965–974.

Monks SA, Leonardson A, Zhu H, Cundiff P, Pietrusiak P, Edwards S *et al.* (2004). Genetic inheritance of gene expression in human cell lines. *Am J Hum Genet* **75**: 1094–1105.

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.

O'Hara RB (2006). Wholesale analysis of genes, traits and microarrays. *Heredity* **97**: 253.

O'Hara RB, Sillanpää MJ (2009). A Review of Bayesian variable selection methods: What, how and which. *Bayesian Analysis* **4**: 85–118.

Perez-Enciso M (2003). Fine mapping of complex trait genes combining pedigree and linkage disequilibrium information: a Bayesian unified framework. *Genetics* **163**: 1497–1510.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**: 904–909.

Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

Purcell S, Sham P, Daly MJ (2005). Parental phenotypes in family-based association analysis. *Am J Hum Genet* **76**: 249–259.

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

Reya T, O'Riordan M, Okamura R, Devaney E, Willert K, Nusse R *et al.* (2000). Wnt signalling regulates B lymphocyte proliferation through a LEF dependent mechanism. *Immunity* **13**: 15–24.

Rubin DB (1976). Inference and missing data. *Biometrika* **63**: 581–592.

Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V *et al.* (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**: 297–302.
Seidensticker M, Behrens J (2000). Biochemical interactions in the Wnt pathway. *Biochim Biophys Acta* **1495**: 168–182.
Setakis E, Stirnadel H, Balding DJ (2006). Logistic regression protects against population structure in genetic association studies. *Genome Res* **16**: 290–296.
Sillanpää MJ, Bhattacharjee M (2005). Bayesian association-based fine mapping in small chromosomal segments. *Genetics* **169**: 427–439.
Sillanpää MJ, Noykova N (2008). Hierarchical modelling of clinical and expression quantitative trait loci. *Heredity* **101**: 271–284.
Spiegelhalter DJ, Thomas A, Best NG (1999). *WinBUGS Version 1.2 User Manual*. MRC Biostatistics Unit: Cambridge, UK.
Thomas DC (1992). Fitting genetic data using Gibbs sampling—an application to nevus counts in 38 Utah kindreds. *Cytogenet Cell Genet* **59**: 228–230.
Thomas DC (2004). *Statistical Methods in Genetic Epidemiology.* Oxford University Press: New York.
Thompson EA, Skolnick MH (1977). Likelihoods on complex pedigrees for quantitative traits. In: Pollack E, Kempthorne O, Bailey Jr TB. (eds). *Proceedings of the International Conference on Quantitative Genetics*. Iowa State University Press: Ames. pp 815–818.
Thornton T, McPeek MS (2007). Case-control association testing with related individuals: A more powerful quasi-likelihood score test. *Am J Hum Genet* **81**: 321–337.
Uimari P, Sillanpää MJ (2001). Bayesian oligogenic analysis of quantitative and qualitative traits in general pedigrees. *Genet Epidemiol* **21**: 224–242.
Visscher PM, Andrew T, Nyholt DR (2008). Genome-wide association studies of quantitative traits with related individuals: little (power) lost but much to be gained. *Eur J Hum Genet* **16**: 387–390.
Voight BF, Pritchard JK (2005). Confounding from cryptic relatedness in case-control association studies. *PLoS Genet* **1**: e32.
Waldmann P (2009). Easy and flexible Bayesian inference of quantitative genetic parameters. *Evolution* (in press).
Weir BS, Anderson AD, Hepler AB (2006). Genetic relatedness analysis: modern data and new challenges. *Nat Rev Genet* **7**: 771–780.
West M, Ginsburg GS, Huang AT, Nevins JR (2006). Embracing the complexity of genomic data for personalized medicine. *Genome Res* **16**: 559–566.
Xu S (2003). Estimating polygenic effects using markers of the entire genome. *Genetics* **163**: 789–801.
Yi N, Xu S (2000). Bayesian mapping of quantitative trait loci under the Identity-by-Descent-based variance component model. *Genetics* **156**: 411–422.
Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF *et al.* (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* **38**: 203–208.
Zhao H (2000). Family-based association studies. *Stat Methods Med Res* **9**: 563–587.
Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C *et al.* (2007). An *Arabidopsis* example of association mapping in structured samples. *PLoS Genet* **3**: e4.

Appendix A

Estimation

To induce sparseness into a model, we are using Jeffreys' prior $p(\sigma^2) \propto 1/\sigma^2$ for variances. This however, is an improper prior (does not integrate to a finite value) which can lead to an improper posterior (for example, Hopert and Casella, 1996; ter Braak *et al*., 2005). As we are using WinBUGS software (Gilks *et al*., 1994; Spiegelhalter *et al*., 1999) for implementation, we need to do some adjustments. WinBUGS uses nonstandard parameterization of distributions in terms of their precision (that is, precision $= \tau = 1/\text{variance}$). We make transformation $\phi = \log(\tau)$ for the precision parameters. Note that the transformation applies equivalently for both variance and precision. Equivalent to the prior $p(\tau) \propto 1/\tau$, the prior for transformed parameter can be derived as $p(\phi) = p(\tau)|\frac{\partial \tau}{\partial \phi}| \propto \frac{1}{\tau}\tau = 1$ (see Gelman *et al*., 2004, p. 65). However, the flat prior $p(\phi)$ is also improper, but when it is restricted to some finite range, it will give us proper prior. We restricted the precision to the range $[\frac{1}{b}, \frac{1}{a}]$, where b is empirical approximation of phenotypic variance and a is very close to zero (10^{-18}). For the precision parameter we also tried a Gamma prior with certain shape parameters which has similar shape as Jeffreys' prior (see Cemgil *et al*., 2007). We found out that such a prior was sensitive to shape parameters and it also easily produced numerical instability ('trap messages') in WinBUGS. Thus, we decided to use a restricted Jeffreys' prior in our examples. The prior for μ is also an improper flat prior. An approximation for that is flat normal density with zero mean and large enough variance.

The prior for the missing data of founders is constructed in the following way: we create two hypothetical extra individuals, which are the parents of all the founders. These artificial individuals are heterozygotes in all their markers. Thus, we can give the same prior $p(m_{i,j} \mid m_{m,j}, m_{f,j})$ for all the individuals regardless of them being founders or non-founders, which allow us to use WinBUGS. In this way, we could keep the data structured by the pedigrees and this procedure is equivalent to the assumption of uniform allele frequencies. We assume that also phenotypic data is missing at random (Rubin, 1976). WinBUGS follows this assumption and thus, the posterior distributions of the parameters are influenced only by the observed records of the outcome variable.

For the infinite polygenic model, we tested both the multivariate normal distribution and the conditional factorization of Lin (1999) and Thomas (1992) as a prior for 'breeding values'. In WinBUGS, our experience confirmed the expectation that both of these methods are practically equally efficient than a block updating of 'breeding values' which maintain well mixing samplers. Faster computational speed favors the use of conditional factorization.