Heredity (2012) 108, 134–146

www.nature.com/hdy

ORIGINAL ARTICLE

Simultaneous estimation of multiple quantitative trait loci and growth curve parameters through hierarchical Bayesian modeling

MJ Sillanpää[1,2], P Pikkuhookana[1], S Abrahamsson[3], T Knürr[1], A Fries[3], E Lerceteau[4], P Waldmann[5,6] and MR García-Gil[3]

[1]*Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland;* [2]*Department of Agricultural Sciences, University of Helsinki, Helsinki, Finland;* [3]*Department of Forest Genetics and Plant Physiology, Swedish University of Agricultural Sciences, Umeå, Sweden;* [4]*Karl-Franzens University of Graz, Institute of Zoology, Graz, Austria;* [5]*Thetastats, Lund, Sweden and* [6]*Division of Livestock Sciences, University of Natural Resources and Applied Life Sciences, Vienna, Austria*

A novel hierarchical quantitative trait locus (QTL) mapping method using a polynomial growth function and a multiple-QTL model (with no dependence in time) in a multitrait framework is presented. The method considers a population-based sample where individuals have been phenotyped (over time) with respect to some dynamic trait and genotyped at a given set of loci. A specific feature of the proposed approach is that, instead of an average functional curve, each individual has its own functional curve. Moreover, each QTL can modify the dynamic characteristics of the trait value of an individual through its influence on one or more growth curve parameters. Apparent advantages of the approach include: (1) assumption of time-independent QTL and environmental effects, (2) alleviating the necessity for an autoregressive covariance structure for residuals and (3) the flexibility to use variable selection methods. As a by-product of the method, heritabilities and genetic correlations can also be estimated for individual growth curve parameters, which are considered as latent traits. For selecting trait-associated loci in the model, we use a modified version of the well-known Bayesian adaptive shrinkage technique. We illustrate our approach by analysing a sub sample of 500 individuals from the simulated QTLMAS 2009 data set, as well as simulation replicates and a real Scots pine (*Pinus sylvestris*) data set, using temporal measurements of height as dynamic trait of interest.
Heredity (2012) **108,** 134–146; doi:10.1038/hdy.2011.56; published online 27 July 2011

Keywords: functional mapping; scots pine; QTL; multitrait; Bayesian model; MCMC

Introduction

Several approaches to mapping quantitative trait loci (QTLs) influencing dynamic traits (that is, traits, of which expression changes over time) have been proposed (see Wu and Lin, 2006 for a review). Even though phenotypes measured at different time points may be controlled by different sets of QTLs, the phenotypic values over time points are generally highly correlated. Thus, the repeated measurement framework has been proposed for QTL analysis of trait measurements over time (Lynch and Walsh, 1998). Alternatively, traits measured at different time points can be treated as separate traits and analysed jointly in a multitrait framework. Here, the efficient parametrisation of multiple trait framework in terms of covariance functions provides a viable approach (Macgregor *et al*., 2005; Lund *et al*., 2008). However, the most common practice is to use some mathematical function to describe dynamic trait behaviour and then map QTLs, which influence this special function using single or multivariate QTL mapping. For example, the logistic growth function (Ma *et al*., 2002; Wu *et al*., 2002, 2003, 2004), as well as polynomial functions (that is, multiple regression model) (Gee *et al*., 2003), and Legendre polynomial (Yang *et al*., 2006; Yang and Xu, 2007) have been proposed for this purpose. The logistic growth function has been justified biologically (West *et al*., 2001). As a criticism, regression using logistic functions fit only growth trajectories that are sigmoidal, that is, monotonically increasing function of time (Yang and Xu, 2007). The Legendre and other orthogonal polynomial fittings (for covariance function) have also been criticised by Pletcher and Geyer (1999). In general, the choice of function should be based on the complexity of the trait trajectory. Even if several methods have been proposed, most of the approaches are limited to single or two-QTL model. Exceptions to this include the methods of Yang and Xu (2007), Min *et al*., 2011, and Heuven and Janss (2010).

Separate age-to-age analysis of QTL for growth has been applied to address questions on QTL stability across time in woody trees (Verhaegen *et al*., 1997; Conner *et al*., 1998; Kaya *et al*., 1999; Lerceteau *et al*., 2001). However, to our knowledge, Ma *et al*. (2004) is the only study where functional QTL mapping has been applied to study growth trajectories in a forest tree

Correspondence: Dr MJ Sillanpää, Department of Mathematics and Statistics, University of Helsinki, PO Box 68, FIN-00014 Helsinki, Finland. E-mail: mjs@rolf.helsinki.fi
Received 18 May 2010; revised 24 May 2011; accepted 6 June 2011; published online 27 July 2011

species. In their work, Ma *et al*. (2004) noted an increased statistical power for QTL detection based on a functional mapping method compared with the alternative QTL time-point analysis.

Functional QTL mapping methods commonly model average curve behaviour with time-specific QTL and environmental effects (Yang and Xu, 2007 and Min *et al*., 2011). Individual-specific variations in these methods are described as deviations from the mean curve behaviour, and these deviations are dependent at neighbouring time points. Exceptions to this common theme are provided by Gee *et al*. (2003) and Heuven and Janss (2010), where all time-dependent behaviour is described with individual-specific curve parameters, which allows hierarchical modeling of QTL effects. These two worlds (hierarchical and non-hierarchical) are conceptually very distinct from each another. In the parametrisation of Gee *et al*. (2003), QTL effects are not time dependent and they affect the shape of the curve rather than having specific effect at particular time points. To describe the functional curve over time, we consider here the approach of Gee *et al*. (2003). As an improvement to their approach (as well as to the approach of Heuven and Janss, 2010), we formulate the whole problem as a single hierarchical model. In our formulation, we simultaneously use multitrait multiple-QTL model and model selection, while estimating functional curve and other model parameters in a Bayesian framework.

Model

Let us consider the population-based sample of individuals where the data sample has been phenotyped with respect to some dynamic trait, and genotyped at a given set of marker loci. Although this represents a typical design in the population-based single-nucleotide polymorphism association studies, the proposed method is directly applicable to a backcross and double haploids in inbred lines, as well as offspring population resulting from outbred line crosses. When handling missing values, we ignore parental (linkage) information completely, so that markers are treated independently. This also means that only marker positions are considered as putative QTL positions. For alternatives, see the subsection dealing with Missing genotype data further on.

Phenotypic model over time points

For each individual i, let us assume that $y_{i,t}$ is the phenotypic value measured at time point t, ($t=1,\ldots,T$). We use the following regression model to describe the phenotypic behaviour over time:

$$y_{i,t} = \beta_{0,i} + \beta_{1,i}a_{t,i} + \beta_{2,i}a_{t,i}^2 + e_{i,t}. \quad (1)$$

Here, $\beta_i = \{\beta_{0,i},\beta_{1,i},\beta_{2,i}\}$ are the curve parameters for individual i, and the errors $e_{i,t}$ are assumed to be independent and normally distributed with mean zero and variance σ_e^2 common for all time points. Because the curve parameters are different across individuals, we pre-specify σ_e^2 to improve parameter identifiability in our hierarchical model (described below). Note that σ_e^2 describes how much measurements at each time-point are allowed to deviate from individual-specific curve (that is, the level of agreement between the data and the growth function). The suitable value for σ_e^2 will depend on the type of the data. For example, for growth data, we have used here $\sigma_e^2 = 0.1$ constantly in our small simulation examples and $\sigma_e^2 = 0.01$ for real data analyses and for QTLMAS 2009 data analysis (the selected value of σ_e^2 should not be too large as this may lead to all the QTL variation being erroneously explained by the residual error). The quantity $a_{t,i}$ is the age of individual i at timepoint t (in calendar time which can be expressed as deviation from the mean age; see Gee *et al*., 2003). For simplicity we consider common time points and same age for all individuals so that $a_{t,i} = t$ for time-points $t=1,\ldots,T$, and all i.

Multiple trait QTL model

We treat the three curve parameters in β_i as three latent traits, and assume that, conditionally on genetic effects, the curve parameters are *a priori* correlated with each other. By making such an assumption, we can hierarchically fit a multitrait QTL model for the curve parameters β_i. For each individual i, let us assume that there are p additively acting marker loci with genotypic values $x_{i,j}$, $j=1,\ldots,p$, coded as 0 or 1 for two homozygotes and the 0.5 for the heterozygote. Given the marker effects ($B_{(k)} = \{B_{1(k)},\ldots,B_{p(k)}\}$, $k=0$, 1, 2), each curve parameter ($\beta_{k,i}$, $k=0$, 1, 2) is modelled as a linear combination (weighted sum) of effects of genotypes $x_{i,j}$ at different loci.

$$\beta_{0,i} = \mu_0 + \sum_{j=1}^{p} I_{j(0)}B_{j(0)}x_{i,j} + \in_{i(0)}. \quad (2)$$

$$\beta_{1,i} = \mu_1 + \sum_{j=1}^{p} I_{j(1)}B_{j(1)}x_{i,j} + \rho_{10}\beta_{0,i} + \in_{i(1)}. \quad (3)$$

$$\beta_{2,i} = \mu_2 + \sum_{j=1}^{p} I_{j(2)}B_{j(2)}x_{i,j} + \rho_{20}\beta_{0,i} + \rho_{21}\beta_{1,i} + \in_{i(2)}. \quad (4)$$

Here, $\mu = \{\mu_0,\mu_1,\mu_2\}$ are the baseline parameters, and $\in_{i(k)}$ are the residuals. Residuals $\in_{i(k)}$ are assumed to be independent and identically normally distributed with mean zero and variance $\sigma^2_{\in(k)}$ Different residual variances are represented in the vector $\sigma^2_{\in} = \{\sigma^2_{\in(0)},\sigma^2_{\in(1)},\sigma^2_{\in(2)}\}$. The autoregressive terms $\rho = \{\rho_{10},\rho_{20},$ and $\rho_{21}\}$ are included in the model to take into account between-trait residual dependencies so that actual residuals can be assumed to be independent. Autoregressive models are usually used to model covariances between different time points in time series data. We use the same principle here to model between trait covariances (cf class D model of Bonney, 1986). Note that even if one-directional dependence is visible in the model, two-way dependence will be induced automatically as $\beta_{0,i}$ and $\beta_{1,i}$ are model parameters rather than observed quantities in the model. Although a model assuming multivariate normally distributed residuals with unstructured covariance matrix would have been a common way to model this phenomenon, we decided to use this autoregressive model on computational grounds.

In the above multiple trait QTL model, we use own indicator variable for each locus and for each trait, $I_{j(k)}$, where $k=0,1$ or 2. Although these indicators provide a natural way to monitor posterior occupancy of QTLs, the real reason for having them in the model is to improve

heritability estimation as shown by Pikkuhookana and Sillanpää (2009). For details, see the subsection dealing with Heritabilities and genetic covariances/correlations.

Although not explicitly shown here, environmental factors like block effects can be easily included as covariates into the QTL model of each trait (2–4). In such models, environmental factors can have different effects on different curve characteristics. Alternatively, before QTL analysis, one can try to first adjust phenotypic data for (constant or time dependent) environmental factors. This means that residuals of the preliminary analysis are taken as phenotypes for consequtive QTL analysis. However, this kind of adjustment is likely to pre-correct the influence on the intercept ($\beta_{0,i}$) only. In general, this kind of pre-correction practice may have many problems (see Martinez *et al*., 2005), which are likely to be more severe for time-dependent covariates.

Hierarchical model

All the models (1–4) presented above are considered simultaneously as parts of a larger hierarchical model. Let us denote the phenotype and marker data as Y and X, respectively. We denote the model parameters jointly as $\theta = \{\beta_1,\ldots,\beta_N,\mu,B_{(0)},B_{(1)},B_{(2)},I_{(0)},I_{(1)},I_{(2)},\rho,\sigma^2_{\in},\tau^2_{(0)},\tau^2_{(1)},\tau^2_{(2)}\}$. Note that this vector includes all the unknown parameters needed in models (1–4). The posterior distribution $P(\theta \mid X,Y)$ is proportional to the joint distribution $P(X,Y,\theta)$ of the data and parameters. This joint distribution can be described as a product of a likelihood $P(Y \mid \theta)$ and the prior $P(\theta \mid X)$, where the likelihood (with a pre-selected value of σ^2_e) is

$$P(Y|\theta) = \prod_{i=1}^{N}\prod_{t=1}^{T}\frac{1}{\sqrt{2\pi\sigma^2_e}}\exp(-\frac{1}{2\sigma^2_e}(y_{i,t} - \beta_{0,i} - \beta_{1,i}a_{t,i} - \beta_{2,i}a^2_{t,i})^2) \quad (5)$$

and the prior is

$$P(\theta|X) = \prod_{k=0}^{2}\prod_{i=1}^{N}P(\beta_{k,i}|\beta_{<k,i}, X, \mu_k, B_{(k)}, I_{(k)}, \rho, \sigma^2_{\in(k)})\times \quad (6)$$

$$\times\prod_{k=0}^{2}\left[P(\mu_k)P(\sigma^2_{\in(k)})\prod_{j=1}^{p}\left[P(I_{j(k)})P(B_{j(k)}|\tau^2_{jk})P(\tau^2_{jk})\right]\right]\times \quad (7)$$

$$\times P(\rho_{10})P(\rho_{20})P(\rho_{21}). \quad (8)$$

Here, the notation $\beta_{<k,i}$ refers to all the preceding terms in β_i that appear before $\beta_{k,i}$. For example, for $\beta_{1,i}$, a preceding term is $\beta_{0,i}$. The functional forms of priors $P(\beta_{k,i} \mid \beta_{<k,i},X,\mu_k,B_{(k)},I_{(k)},\rho,\sigma^2_{\in(k)})$ are normal densities of the residuals of models (2–4) with mean zero and variance $\sigma^2_{\in(k)}$ (see Sillanpää and Arjas, 1998). For the intercept, this is

$$P(\beta_{0,i}|\beta_{<0,i}, X, \mu_0, B_{(0)}, I_{(0)}, \rho, \sigma^2_{\in(0)}) = \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2_{\in(0)}}}\exp(-\frac{1}{2\sigma^2_{\in(0)}}(\beta_{0,i} - \mu_0 - \sum_{j=1}^{p}I_{j(0)}B_{j(0)}x_{i,j})^2) \quad (9)$$

Each individual prior in $P(\sigma^2_{\in(k)})$ is assumed to be an Inverse-Gamma (0.001, 0.001) and each of $P(\mu_k)$, $P(\rho_{10})$,$P(\rho_{20})$ and $P(\rho_{21})$ to be $N(0, 100)$. The Inverse-Gamma distribution supports values in positive range and the above normal distribution is rather flat. Therefore, they present practical priors applicable for many data sets without normalisation. The priors $P(I_{j(k)})$, $P(B_{j(k)} \mid \tau^2_{jk})$ and $P(\tau^2_{jk})$ are covered in next section.

Model selection

Variable selection, selecting a specific set of trait loci contributing to each of the curve parameters, is here performed using the Bayesian adaptive shrinkage presented in Xu (2003). The adaptive shrinkage of Xu (2003) was found to perform well in comparison with other methods (O'Hara and Sillanpää, 2009). Following Xu (2003), a hierarchical prior

$$P(B_{(0)}, \tau^2_{(0)}) = P(B_{(0)}|\tau^2_{(0)})P(\tau^2_{(0)}) = \prod_j\left[P(B_{j(0)}|\tau^2_{j0})P(\tau^2_{j0})\right]$$

is assumed for coefficients so that $B_{j(0)} \sim N(0,\tau^2_{j0})$ and $P(\tau^2_{j0}) \propto 1/\tau^2_{j0}$. It has been shown earlier that this formulation is mathematically equivalent to assuming Student's t-distribution for $B_{j(0)}$ (Yi and Xu, 2008), which may induce a sparse model representation (Figueiredo, 2003; Xu, 2003; Hoti and Sillanpää, 2006). Similar assumptions are also made in

$$P(B_{(1)}, \tau^2_{(1)}) = \prod_j P(B_{j(1)}|\tau^2_{j1})P(\tau^2_{j1})$$

and in

$$P(B_{(2)}, \tau^2_{(2)}) = \prod_j P(B_{j(2)}|\tau^2_{j2})P(\tau^2_{j2}).$$

The benefit of the adaptive shrinkage is that no tuning is needed, but one cannot make any prior assumptions from the degree of sparseness either.

As in Pikkuhookana and Sillanpää, (2009), there is another source of sparseness in our model, given by the indicator variables. In principle, the degree of sparseness can be controlled by specifying a small prior probability to include each marker into the model. Thus, as prior $P(I_{j(k)})$ for each marker j and each trait k (referring to one of the curve characteristics $k = 0,1,$ or 2), we can assume a Bernoulli distribution with fixed parameter $s = P(I_{j(k)} = 1) = \frac{1}{p}$. However, we know from our earlier experience (Pikkuhookana and Sillanpää, 2009) that the shrinkage prior tends to dominate marker selection, so that this latter Bernoulli prior only has a modest influence on the degree of sparseness.

Missing genotype data

So far, we have implicitly assumed that QTLs are placed exactly at marker points. We assume that there may be a small amount of missing values among the genotypes $x_{i,j}$ of the data sample. In such case, the right hand side of the equation $P(X,Y,\theta) = P(Y \mid \theta)P(\theta \mid X)$, presented in the Hierarchical model above, should also include prior for $P(X) = \Pi_i\Pi_j P(x_{i,j})$. For this we have simply used Bernoulli prior $P(x_{i,j})$ where all genotypic values are considered to be equally likely. In case of a population-based association studies, the use of known or estimated allele frequencies and assuming that genotypes occur in Hardy–Weinberg proportions may provide a more

informative alternative to be used here. Although not considered here, it is possible to build up more efficient missing data models to account for dominant markers, linkage information and/or linkage disequilibrium information potentially available in the data sample. The use of linkage information necessitates the presence of haplotype information and/or multiple generations of pedigree data. The requirement for the use of linkage disequilibrium information is a dense set of markers and known physical or genetic marker distances. Missing marker (or QTL) genotype at arbitrary map positions can be predicted as pseudomarkers based on Mendelian segregation and marker distances (see Sillanpää and Arjas, 1999; Servin and Stephens, 2007 for details) and/or on linkage disequilibrium (Druet and Georges, 2010; Marchini and Howie, 2010).

Markov chain Monte Carlo estimation and posterior summaries

We apply Markov chain Monte Carlo (MCMC) estimation to draw dependent samples from the joint posterior distribution of the unknowns (Robert and Casella, 2004). As an output from adaptive shrinkage, one obtains posterior estimated effect size at each considered position along the genome. Instead of monitoring the estimated effects or their posterior functions (Xu, 2003; Hoti and Sillanpää, 2006), one can take posterior expectations of indicator variables (see equations 2–4) to obtain estimates for posterior occupancy probabilities $P(I_{j(k)}=1\,|\,\text{data})$. This is calculated simply as a proportion of MCMC rounds, in which the focal indicator is one. The posterior $P(I_{j(k)}=1\,|\,\text{data})$ provides a natural model-averaged measure of evidence for a strength of phenotype–genotype association at locus j. Note that we obtain separate set of estimated posterior occupancy probabilities for each of the three curve parameters. For small QTL probabilities, it may be more meaningful to present them as Bayes factors (Kass and Raftery, 1995; Yi *et al*., 2007): $BF_{j(k)}=\frac{P(I_{j(k)}=1|\text{data})/P(I_{j(k)}=0|\text{data})}{P(I_{j(k)}=1)/P(I_{j(k)}=0)}$, which measures evidence for inclusion against exclusion of a locus. Although values of BF in range 1 to 3 are 'not worth more than bare mention', values in interval (3, 10) represents 'substantial' evidence (Jeffreys, 1961). Alternatively, a level of $2ln(BF)=2.1$ has been suggested for declaring statistical significance (Kass and Raftery, 1995).

Heritabilities and genetic covariances/correlations

As polynomial curve parameters (intercept, slope and quadratic terms) are treated as latent traits in our model, we can estimate posterior heritabilities for each of them based on marker data. Like the curve parameters, these heritabilities are constant over time points. Using the QTL models (2–4), (narrow sense) heritabilities for curve parameters $\{\beta_0,\beta_1,\beta_2\}$, can be estimated when there are no environmental effects in the QTL models. For the intercept, $h^2_{y0}\approx\frac{1}{M}\sum_{m=1}^{M}\frac{\hat{\sigma}^2_{y0}(m)-\sigma^2_{\epsilon(0)}(m)}{\hat{\sigma}^2_{y0}(m)}$, where $\hat{\sigma}^2_{y0}(m)$ is the empirical phenotypic variance of β_0 at MCMC round m, which can be estimated using sample variance of the curve parameter $\beta_{0,i}(m)$ values at MCMC round m. Here, M is total number of MCMC rounds, $\sigma^2_{\epsilon(0)}(m)$ is residual variance for an intercept in MCMC round m. Heritabilities for the slope $h^2_{y1}\approx\frac{1}{M}\sum_{m=1}^{M}\frac{\hat{\sigma}^2_{y1}(m)-\sigma^2_{\epsilon(1)}(m)}{\hat{\sigma}^2_{y1}(m)}$ and the quadratic term $h^2_{y2}\approx\frac{1}{M}\sum_{m=1}^{M}\frac{\hat{\sigma}^2_{y2}(m)-\sigma^2_{\epsilon(2)}(m)}{\hat{\sigma}^2_{y2}(m)}$ are based on terms from QTL models (3) and (4), respectively. In the presence of environmental effects, $h^2_{y0}\approx\frac{1}{M}\sum_{m=1}^{M}\frac{Var[\sum_j I_{j(0)}B_{j(0)}x_{i,j}](m)}{\hat{\sigma}^2_{y0}(m)}$ where the numerator is the empirical variance of the predictor of the QTL model (2) in MCMC round m. In general, as was found by Pikkuhookana and Sillanpää (2009), indicator variables in QTL models (2–4) improve heritability estimation. Otherwise, the cumulative sum of markers with spurious effects tends to introduce some noise to the predictions.

The use of multitrait QTL model makes it possible to also estimate genetic covariances (and correlations) between polynomial curve parameters. The genetic covariance between the intercept and the slope is estimated as $\sigma_{g10}\approx\frac{1}{M}\sum_{m=1}^{M}(\hat{\sigma}_{y10}(m)-\rho_{10}(m)\hat{\sigma}_{y10}(m))$ and the genetic correlation as $r_{10}\approx\frac{1}{M}\sum_{m=1}^{M}\frac{\sigma_{g10}}{\hat{\sigma}_{y1}(m)\hat{\sigma}_{y0}(m)}$.

Here $\hat{\sigma}_{y10}(m)=[\sum_{i=1}^{N}\beta_{1,i}(m)\beta_{0,i}(m)-\frac{(\sum_i\beta_{1,i}(m))(\sum_i\beta_{0,i}(m))}{N}]/N-1$ is the empirical covariance between the intercept and the slope at MCMC round m. The term $\rho_{10}(m)\hat{\sigma}_{y0}(m)$ represents the residual covariance in MCMC round m. The genetic covariances (σ_{g21} and σ_{g20}) and genetic correlations (r_{21} and r_{20}) for other parameters are calculated using the same principle.

Example analyses

Simulated data from QTLMAS 2009

We used public-simulated time-course data from QTLMAS 2009 workshop (Coster *et al*., 2010), which has been previously analysed in several other publications including Heuven and Janss (2010). The data set consists the growth curve measurements at five consecutive time points and 453 markers (within five chromosomes) measured from 2025 individuals. There were certain family structures present among the individuals in the data. There were altogether 18 QTLs influencing the growth curve phenotype among which three QTLs had five times larger effects on the trait than the rest of the QTLs. A map is available at http://www.qtlmas2009.wur.nl/UK/Dataset/ where one can find the marker ID and positions. The individual growth curves were simulated based on a logistic growth function, which is different from our polynomial growth function (equation 1). Thus, the logistic growth data can be seen as a test of the robustness of our method. We took a sub sample of 500 individuals, selected randomly within each family (but with equal contribution from all full-sib families). Only 50 families with phenotype data were used, which resulted in 10 individuals per family.

Real data on scots pine (*Pinus sylvestris*)

We genotyped a set of 160 AFLPs on 250 individuals from a full-sib family of Scots pine that was established in 1988. The parents of the full-sib cross are part of the

Swedish breeding population; both parents come from northern Sweden (AC3065 latitude 6508′ and Y3088 latitude 6409′).

Total DNA was extracted from vegetative buds. The buds where pealed, dried and grinded. The DNA extraction was made using the CTAB method. The AFLP markers were produced according to Vos *et al*. (1995). The following 15 primer enzyme combinations were used E-act/M-cctg, E-act/M-cccg, E-act/M-ccgc, E-act/M-ccgg, E-act/M-ccag, E-acg/M-cctg, E-acg/M-cccg, E-acg/M-ccgc, E-acg/M-ccgg, E-acg/M-ccag, E-aca/M-cctg, E-aca/M-cccg, E-aca/M-ccgc, E-aca/M-ccgg and E-aca/M-ccag.

The amplified fragments were sent to the DNA facility at Iowa State University, USA and run on ABI3100 Genetic Analyzer. The mapping data were analysed with GeneMarker v1.6 (SoftGenetics, State College, PA, USA).

The height measurements were carried out with a measuring stick of telescope type from the ground to the terminal bud. The height was repeatedly measured 11 times between the years 1996 and 2007. The phenotype measurements from 1996 to 1999 have already been used for QTL analysis and published in Lerceteau *et al*. (2001). After more close inspection of temporal measurements, we decided to exclude 14 individuals from the collected data because those individuals showed negative enrichment of height in some of the consecutive time points because of some damage in the apical shoot due to wind or snow. Thus, our final data set contained 236 individuals.

Simulated pine data replicates

Latent trait phenotypes: We took above real Scots pine data (236 individuals, 160 AFLP markers; $x_{i,j}$, $i=1,\ldots,236$ $j=1,\ldots,160$) as starting point for our simulation. First with equal probabilities we completed (by sampling once) all the missing genotypes so that there was no missing genotypes in simulated data. For each individual, we simulated 10 replicates of a vector β_i containing a new set of latent trait phenotypes from the modified versions of the QTL models (2–4) by setting $\rho_{10}=\rho_{20}=\rho_{21}=0$ and assuming that the residual vector $\in_i=(\in_{i(0)}, \in_{i(1)}, \in_{i(2)})$ is drawn from a tridimensional normal distribution, $MVN(\tilde{0},\Sigma)$, with a mean vector $\tilde{0}=(0,0,0)^t$ and a covariance matrix Σ specifying between-trait residual dependencies. For all the replicates, three QTLs (at loci 18, 32 and 95) with average joint heritability of 0.39 (replicates varied in range (0.36–0.47)) were simulated for the 1st latent trait–intercept, four QTLs (at loci 32, 74, 135 and 144) with average heritability of 0.66 (replicates in (0.62–0.71)) for the 2nd latent trait-slope and two QTLs (at loci 9, 104) with average heritability of 0.50 (replicates in (0.43–0.56)) for the 3rd latent trait—quadratic term. Here, indicators $I_{j(k)}, j=1,\ldots,160$, $k=0, 1, 2$ were set to one for QTLs and to zero for non-QTLs. The contents of Σ and the other QTL-model parameters used in our simulations are described in Tables 1 and 2.

Functional trait phenotypes: Given the replicated values of curve parameters (that is, latent traits) above, we simulated 10 replicates of phenotypic measurements at 11 consecutive time points for each individual. For this, we used the modified version of the model (1) with time-specific residual variances $\{\sigma^2_1,\ldots,\sigma^2_{11}\}=\{2,3,5,4,3,5,3,2,3,5,4\}$.

Table 1 Analysis of 10 simulated data replicates

Latent trait	*Location*	*Simulated effect*	*E(effect × I_j | data)*	*P($I_j=1$ | data)*
Intercept	18	1/2	0.0002	0.006
	32	3/2	0.152 (0, 0.858)	0.109
	65	0	0.076	0.011
	75	0	−0.090	0.061
	78	0	0.077	0.058
	95	−1/2	−0.0008	0.006
	99	0	0.021	0.022
Slope	32	2	2.222 (1.878, 2.562)	1.000
	74	−1/2	−0.036	0.056
	135	−1	−0.867 (−1.253, −0.445)	0.907
	144	3/2	1.440 (1.073, 1.792)	1.000
Quadratic	9	−1/2	−0.000009	0.359
	76	0	0.003	0.014
	104	2	1.971 (1.709, 2.232)	1.000

$$\sum = \begin{pmatrix} 1 & 0.5 & 0.3 \\ 0.5 & 1 & 0.2 \\ 0.3 & 0.2 & 1 \end{pmatrix}$$

The map locations and the simulated and estimated phenotypic (additive genetic) effects of the trait loci of three latent traits (intercept, slope and quadratic term), as well as their posterior occupancy probabilities $P(I_j=1 \mid \text{data})$. The marker effect is estimated as $E(\text{effect} \times I_j \mid \text{data})$ and the corresponding 95% credible interval is shown for the estimates, of which absolute value is larger than 0.1 . The posterior estimates are averaged over analyses of 10 replicated data sets. The values used in the residual covariance matrix Σ on simulations is also shown

Table 2 Analysis of 10 simulated data replicates

Parameter	*Simulated value*	*Posterior mean estimate*	*95% credible region*
h^2_{y0}	0.39	0.009	(−0.195, 0.179)
h^2_{y1}	0.66	0.544	(0.443, 0.627)
h^2_{y2}	0.50	0.499	(0.396, 0.584)
N_{y0}	3	1.1826	(0.0, 3.3)
N_{y1}	4	3.8579	(2.9, 5.9)
N_{y2}	2	2.2561	(1.3, 4.5)
σ_{g10}	1.48	−0.246	(−0.447, −0.044)
σ_{g21}	0.13	0.021	(−0.116, 0.156)
σ_{g20}	0.26	0.226	(0.073, 0.379)
r_{10}	0.61	−0.046	(−0.084, −0.007)
r_{21}	0.05	0.009	(−0.042, 0.060)
r_{20}	0.13	0.054	(0.0163, 0.092)
$\sigma^2_{\in(0)}$	1.00	7.798	(6.428, 9.450)
$\sigma^2_{\in(1)}$	1.00	1.578	(1.291, 1.925)
$\sigma^2_{\in(2)}$	1.00	1.039	(0.861, 1.250)
μ_0	3.00	3.563	(2.878, 4.106)
μ_1	4.00	4.242	(3.838, 4.686)
μ_2	5.00	4.396	(3.922, 4.870)
$\rho_{10}\times\hat{\sigma}_{y0}$	0.5	−0.536	(−0.288, −0.359)
$\rho_{21}\times\hat{\sigma}_{y1}$	0.2	0.176	(0.040, 0.311)
$\rho_{20}\times\hat{\sigma}_{y0}$	0.3	0.093	(−0.042, 0.229)

The simulated values and the posterior estimates (mean and 95% credible region) of heritabilities, the number of QTLs, genetic covariances, genetic correlations, residual variances, the baselines and pairwise residual covariances (cf.Table 1) for the three latent traits (intercept *y*0, slope *y*1 and quadratic term *y*2). The posterior estimates are averaged over analyses of 10 replicated data sets.

These time-specific residual variances describe how much individual phenotypic measurements (at each time point) are allowed to deviate from individual-specific functional curve. Eventually this process produced 10 data replicates.

Creation of data sets with missing phenotypes: As a goodness-of-fit test for the model, we wanted to study also the robustness of our method for increased number of missing phenotype measurements in time points. Thus, for every other time point, we introduced missing entries by deleting $\sim 50\%$ of the phenotypes randomly. Note that individuals with missing phenotypes at consecutive time points are not neccessarily the same. The same treatment was carried out for a real Pine data set and one additional simulation replicate. After this treatment, we had 13 different Pine data sets: an original real data set, 10 simulation replicates, and one real and one simulated data set with increased missingness.

Analyses

In the following, we introduce results from five different analyses. First, we present QTL analysis of simulated QTLMAS 2009 data set. Then, we cover QTL analysis of 10 simulation replicates and prediction of unobserved phenotypes for additional simulated Pine data set. Finally, we show results from QTL analysis and phenotype predictions with real Pine data. In real data analyses, we included environmental block effect (of four blocks) to each of the QTL models (2–4) and assumed that block effects in each model are independently normally distributed with common block variance. For three block variances, we assumed Inverse-Gamma (0.01, 0.01) priors.

For implementation and parameter estimation, we used WinBUGS 1.4.3 software (Spiegelhalter *et al*., 2005). We assumed prior $s = P(I_{j(k)} = 1) = \frac{1}{453}$ for each locus j and for each latent trait k in QTLMAS 2009 data analysis and for each j and k in simulated and real Pine data analyses. As WinBUGS does not allow the use of improper priors such as $P(\tau^2_{jk}) \propto 1/\tau^2_{jk}$, we used its finite approximation (for details, see Pikkuhookana and Sillanpää, 2009). For each data replicate, we ran one chain for 30 000 MCMC iterations, discarding 5000 initial samples as burn-in and thinning the remainder to each 10th sample (that is, storing every 10th sample). This resulted in 2500 samples to be used in estimating the posterior for each data replicate. For prediction of unobserved phenotypes in simulated and real data, as well as the real Pine data QTL-analysis, we ran one chain for 50 000 MCMC iterations and used a burn-in period of 5000 samples and a thinning of 5. This resulted in 9000 MCMC samples. The MCMC sample paths of several different parameters were visually inspected based on some prior runs. The running time was practically the same in phenotype prediction analyses and in real Pine QTL analysis, being about 114 h for the whole analysis on an Intel Core 2 with 1.86 GHz and 1.94 GB of RAM. On the same computer, running through 10 simulation replicates took about 30 days. For the QTLMAS 2009 data analysis, we ran 10 000 MCMC iterations by omitting 6000 initial samples as burn-in and had no thinning.

In the missing data analyses, the prediction accuracy (between true and predicted phenotypes) was assessed at each time point (with increased missingness) by monitoring posterior distributions of relative and absolute prediction error and linear correlation between true and predicted phenotypes. Calculations of these quantities were based on posterior predictive distributions for individuals with missing phenotypes. As stated in Lee *et al*. (2008), this kind of analysis gives information also about the accuracy of this method in estimating genomic breeding values (Meuwissen *et al*., 2001; Piyasatian *et al*., 2007; Lorenzana and Bernando, 2009; Heffner *et al*., 2009).

Results

Simulated QTLMAS 2009 data set

QTL identification: The loci showing elevated posterior occupancy probabilities in the three latent traits are shown in Table 3. The occupancy probabilities of all the other loci were lower than 0.01. The posterior estimated heritabilities for the three latent traits are shown in Table 4. As expected, because different growth functions were used during simulation and analysis, the QTLs simulated for one trait seem to be 'scattered' among all three traits in the analysis. The same phenomenon is visible also in the estimated heritabilities.

The three major QTLs in the data set (at 36, 51 and 78) were in chromosome 1 with map locations 0.4245, 0.5425 and 0.8765 (see Figure 3 in Coster *et al*., 2010). Locus 36 or 35, adjacent to the first major QTL (at 36), showed QTL-occupancy probability of 1.0 in all three latent traits. The QTL probabilities of 0.058 and 0.72 were found for loci 38 and 37 in slope and quadratic term, respectively. These positions are evidently more close to the first major QTL (at 36) but the second major QTL (at 51) is not more than 10 cM away from them. The locus 81, which is close to the third expected QTL at position 78, acquired QTL probability of 1.0 for the slope.

Table 3 QTLMAS 2009 data analysis

Latent trait	*Locus*	*Closest simulated QTL*	*E(effect × I_j | data)*	*P(I_j = 1 | data)*
Intercept	36 (0.4153)	(0.4245) 36–37	1.528	1.0
	98 (1.0359)	(1.0455) 98–99	0.005	0.015
	137 (1.4829)	(1.4889) 138–139	0.177	0.59
	173 (1.8574)	(1.8864) 174–175	−0.017	0.034
	218 (2.2707)	(2.2622) 216–217	−0.004	0.012
	232 (2.4005)	(2.5609) 243–244	0.2022	0.17
	288 (3.048)	(3.0962) 293–294	−0.295	0.68
	408 (4.5353)	(4.5971) 411–412	0.019	0.055
	421 (4.6635)	(4.7719) 432–433	0.013	0.022
Slope	36 (0.4153)	(0.4245) 36–37	0.438	1.0
	38 (0.4472)	*(0.5425) 51–52	0.005	0.058
	81 (0.9137)	(0.8765) 77–78	0.295	1.0
	240 (2.5252)	(2.5609) 243–244	0.001	0.017
	360 (3.8701)	(3.8639) 358–359	0.141	1.0
Quadratic	35 (0.4029)	(0.4245) 36–37	−0.130	1.0
	37 (0.4447)	*(0.5425) 51–52	0.145	0.72
	118 (1.3058)	(1.3302) 118–119	−0.001	0.014
	134 (1.4743)	(1.4889) 138–139	0.001	0.017
	140 (1.5242)	(1.4889) 138–139	0.001	0.012
	222 (2.3108)	(2.2622) 216–217	−0.001	0.018
	223 (2.318)	(2.2622) 216–217	0.001	0.010
	314 (3.3746)	(3.3652) 313–314	−0.001	0.014

The three latent traits are listed in the first column. The second column 'Locus' refers to the marker loci (with their map locations in parenthesis), which showed non-negligible signals in QTL analysis. The next column refers to the map position of true QTLs (in parenthesis), followed by their two flanking markers. The symbol * indicates the position of the second closest major QTL. The posterior means of the marker effects (viz. $B_{j(k)} \times I_{j(k)}$ for latent trait k) and the posterior occupancy probabilities $P(I_j = 1 \mid \text{data})$ for detected loci j are given in the last two columns.

Table 4 QTLMAS 2009 data analysis

Parameter	*Simulated value*	*Posterior mean estimate*	*95% credible region*
h^2_{y0}	0.50	0.24	(0.13, 0.34)
h^2_{y1}	0.50	0.99	(0.984, 0.995)
h^2_{y2}	0.50	0.75	(0.71, 0.79)

The posterior estimates (mean and 95% credible region) of heritabilities of three latent traits (intercept $y0$, slope $y1$ and quadratic term $y2$). Note that simulated values correspond to the heritabilities simulated under different growth function.

Table 5 QTLMAS 2009 data analysis

Parameter	β_0	β_2	β_3
par_1	0.46	−0.37	0.61
par_2	0.42	−0.35	0.29
par_3	−0.29	0.41	−0.13

The pairwise correlation calculated between the three parameters of the logistic growth function (par_1, par_2, par_3) and the three posterior mean-estimated parameters of the polynomial function (β_0,β_1,β_2).

Markers near the four minor QTLs (98, 118, 138 and 174) in chromosome 2 also obtained some support in the analysis. The markers 98, 137 and 173 obtained elevated signals in the intercept and markers 118, 134 and 140 in the quadratic term. All of them are extremely close to the one of four simulated minor QTLs in chromosome 2. Generally, the level of support was not strong, expect for locus 137 where the posterior QTL probability was 0.59.

The markers near two minor QTLs (217 and 243) out of four simulated minor QTLs in chromosome 3 got support in the analysis, but the level of support was generally quite small. Such putative QTL positions were the markers 218 and 232 (with QTL probabilities 0.012 and 0.17, respectively) for intercept, the marker 240 (with QTL probability 0.017) for slope, and the markers 222 and 223 (with QTL probabilities 0.018 and 0.01, respectively) for quadratic term.

The markers near three minor QTLs (293, 314 and 358) out of four simulated minor QTLs in chromosome 4 got support in the analysis. The QTL probabilities for loci 288, 314 and 360 were 0.68, 0.014 and 1.0, respectively. The simulated minor QTL, which was not found in the analysis had small effect size.

The markers near two minor QTLs (411 and 432) out of three simulated minor QTLs in chromosome 5 got support in the analysis. The loci 408 and 421 had QTL probabilities 0.055 and 0.022 in intercept, respectively. The simulated minor QTL, which was not found in the analysis, had a slightly larger effect size than the two others.

To better understand the 'scattering' of QTLs among latent traits, we also calculated the pairwise correlation between the orginal simulation parameters of the logistic growth function and the posterior estimated values of the three latent traits (see Table 5). As can be seen in the table, generally these correlations are moderate except the correlation of 0.61 between logistic curve parameter 1 and the quadratic term.

Heuven and Janss (2010) analysed a sub sample of 1000 individuals from the same QTLMAS 2009 data set by using the growth function assumed in the data simulation process. Because of this, their heritability and QTL position estimates showed consistency among the latent trait parameters. To compare the genomic positions of the found QTLs roughly (without caring about which trait each QTL contributes to or how strong QTL signals were obtained), it is fair to say that comparable set of QTL positions were identified in our analyses with the data sample, which was only half of their sample size.

Simulated Pine data sets

QTL identification: To assess empirical power of our method in simulated Pine data, the estimated posterior occupancy probabilities (averaged over 10 data replicates) for true and false QTLs in the three latent traits are shown in Table 1. Generally, the major QTLs with effect size of at least one (in absolute value) were correctly found in all cases, whereas the QTLs with effect size 1/2 were correctly identified only once and was unidentified three times. For the intercept, the highest QTL-occupancy probability 0.1 was found at correct major QTL with large effect (at locus 32), but the signal was very low. The weak QTL probabilities were found at loci 65, 75, 78 and 99, which were all false positives. All the other QTL probabilities were smaller than 0.01 and no signals (QTL probabilities 0.006 and 0.006) were found at minor QTLs (18 and 95). For the slope, the high posterior occupancy probability (between 0.9 and 1.0) was obtained for three true QTLs with large effects (at loci 32, 135 and 144) and 0.06 for the fourth minor QTL. Practically, there were no false positives in slope because all the other positions had QTL probabilities, which were smaller than 0.01. Note that correctly identified QTL at locus 32 was pleiotropic and had large effects on both intercept and slope. For the quadratic term, all simulated QTLs were correctly identified so that the posterior occupancy probability 1.0 was obtained for the major QTL at locus 104 and probability of 0.36 for minor QTL at locus nine. The weak signal (QTL probability 0.014) was found at locus 76, which was false positive and all the other loci had QTL probability that was smaller than 0.01. It is worth emphasizing here that putting the QTL probability threshold to 0.1 would result in the elimination of all the false positives in these data.

Estimation of the model parameters: As the indicator and effect size always appear together as a pair in the models (2–4), the two quantities are obviously confounded in their estimates. Thus, the QTL-effect estimates are presented only in the form of the product in Table 1. Generally, these posterior means of the estimated QTL-effects (averaged over replicates) are clearly closer to their true simulated values when the QTL-occupancy probability is high, and are constantly small when the corresponding QTL probability is low. The exception to this is the minor QTL in quadratic term where QTL probability was 0.36 but the effect size was practically zero. The simulated and estimated values for several other model parameters are presented in Table 2. The posterior mean estimate of the heritability for the intercept was 0.01, which was much lower than the true simulated mean value of 0.39. Here the 95 % credible interval does not contain the true value. These estimates are biased probably because the support for the major

QTL was weak and two minor QTLs were unidentified in the QTL analysis. For the slope, we obtained an underestimated posterior mean heritability of 0.54 while the true value was 0.66. Here the 95% credible interval averaged over replicates does not contain the true simulated value because one minor QTL was unidentified in the QTL analysis. For the quadratic term, heritability was accurately estimated with posterior mean 0.50, which coincides with the true value. Moreover, the 95% credible interval was rather narrow, indicating that both simulated QTLs for the quadratic term were correctly identified. On the other hand, the number of QTLs is accurately estimated for the slope and the quadratic term while that for the intercept is badly underestimated.

Although the posterior means of the parameter estimates connected to the slope and quadratic term are generally very close to their true simulated values, the estimates connected to the intercept are much more biased (for example, residual variance). An exception to this general trend comes from the baseline parameters, which coincide closely with their true values in all cases. Similarly, the posterior mean of genetic covariance 0.23 and of genetic correlation 0.05 between an intercept and a quadratic term are not very far from their true values (0.26 and 0.13, respectively), while the corresponding parameters between a slope and a quadratic term are more biased. To compare the estimate of ρ_{10} to the simulated residual covariance, we first have to multiply it with the standard deviation of latent trait $\hat{\sigma}_{y0}$, which gives $\hat{\sigma}_{10} = \rho_{10} \times \hat{\sigma}_{y0} \approx -0.5$, which is not close to 0.5. Similar reasoning yields estimates of $\hat{\sigma}_{21} \approx 0.18$ and $\hat{\sigma}_{20} \approx 0.1$, which are closer to values of 0.2 and 0.3.

Prediction of unobserved phenotypes: For simulated data, the values of correlation coefficients between the true and predicted phenotypes (their posterior means) were calculated at five different time points with increased amount of missing phenotypes. These correlations were almost one in all cases (0.992, 0.998, 1.0, 1.0 and 1.0 for five time points), which indicates that our method was able to correctly predict the original ordering of the unobserved phenotypes. For simulated data, the boxplots in Figure 1 present the relative errors (top) and absolute errors (bottom) of the predicted phenotypes at five time points with increased amount of missing phenotypes. These quantities are calculated from posterior predictive distributions of unobserved phenotypes. The reason why relative errors are decreasing as function of time is the fact that our simulated growth phenotype is systematically increasing with time. This means that error values on the right are systematically divided by larger (true) values. In this case, absolute error is providing better indication of phenotype prediction accuracy. At each measurement point, the mean of the absolute error stays in the vicinity of zero and one cannot see the systematic trend of making larger absolute errors, while the actual predicted values increase from left to right.

Real data on Scots pine

QTL identification: From the real data analysis, the estimated posterior occupancy probabilities and effect sizes of QTLs in the three latent traits are shown in Table 6. Even though these QTL probabilities are

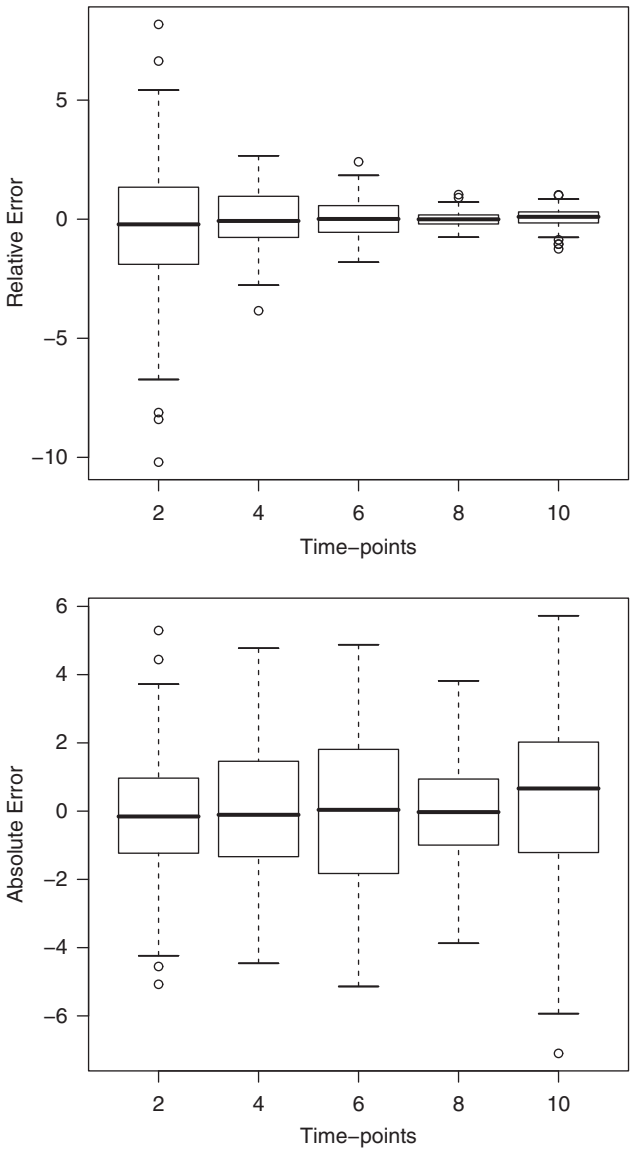

Figure 1 The prediction accuracy of the phenotypes in the simulated data. Boxplots of relative errors (top) and absolute errors (bottom) for predicted phenotypes at five time points, which had increased missingness. Errors are shown on the y-axis and the time-points on the x-axis. The absolute error is calculated as difference between predicted (posterior mean) and true phenotype and relative error is obtained as 100 times absolute error divided by the absolute value of the true phenotype.

Table 6 Scots pine data analysis

Latent trait	*Location*	E(*effect* × I_j | *data*)	P(I_j = 1 | *data*)	*Bayes Factor*
Intercept	156	0.017	0.009	1.50
Slope	21	0.018	0.012	1.90
Quadratic	38	−0.0006	0.012	1.88
	97	0.001	0.013	2.17

The locations and the estimated phenotypic (additive genetic) effects of the trait loci of three latent traits (intercept, slope and quadratic term), as well as their posterior occupancy probabilities $P(I_j = 1 \mid \text{data})$ and the corresponding Bayes factors. The effect size is estimated as $E(\text{effect} \times I_j \mid \text{data})$.

generally low, our suggested loci show clearly elevated signals compared with the general level of that in other positions. However, based on our replicated simulation analysis, our power here may be rather weak. As it is hard to judge small QTL probabilities, we decided to present also the Bayes factor (BF) as BF scales the corresponding marker evidence with respect to the prior probability. For the intercept, the highest QTL probability 0.009 (BF 1.50) was found for locus 156 (*act/ccgg_*433), while the other QTL probabilities were all smaller than 0.0085 (BF $<$1.335). However, at QTL threshold level 0.01, which is rather low, one can conclude that no QTLs were found for intercept. For the slope, the highest QTL probability 0.012 (BF 1.90) occurs at locus 21 (*aca/ccgc_*194) and all the other QTL probabilities were smaller than 0.009 (BF $<$1.391). For the quadratic term, there were two putative QTLs (at loci 38 and 97; *aca/ccgg_*277 and *acg/ccgc_*71) with QTL probabilities 0.012 and 0.013 (BFs 1.88 and 2.17, respectively). The other QTL probabilities were smaller than 0.008 (BF $<$1.157). Based on the general BF categories suggested by Jeffreys (1961), these evidence are from class of 'not worth more than a bare mention'. However, one should keep in mind that there are two sources of shrinkage in our QTL models, where the indicators had milder influence on the overall shrinkage than the effect coefficients (Pikkuhookana and Sillanpää, 2009). Thus, in the presence of small sample size, it might be fair to conclude that the BFs presented here are kind of 'lower bounds of their true values' and should be interpreted in the light of much smaller prior inclusion probability. However, even if influence of this 'double shrinkage' would be modest, it is likely that these findings are still rather weak.

Estimation of the heritabilities and other parameters: The posterior estimates for different model parameters including latent trait heritabilities are shown in Table 7. The heritabilities are small for all latent traits, which supports the fact that genetic variation is generally low.

Prediction of unobserved phenotypes: For real data, the value of linear correlations coefficient between the true and predicted phenotypes (their posterior means) were calculated at five different time points with increased amount of missing phenotypes. As in simulated data, the correlation coefficients here were also extremely high in all cases (0.993, 0.984, 0.995, 0.995 and 0.980 for five time points), which indicates that our method was able to correctly predict the original ordering of the unobserved phenotypes. It is likely that environmental block effects are at least partly responsible for these predictions, because the estimated QTL probabilities and heritabilities in these data set were so small. For real data, the boxplots in Figure 2 present the relative errors (top) and absolute errors (bottom) of predicted phenotypes at five time points with increased amount of missing phenotypes. It is clear that the first time point seems to suffer from some bias, which may reflect a disagreement between the polynomial function and the data or difficulties in mapping QTLs for the intercept.

Table 7 Scots pine data analysis

Parameter	*Posterior mean estimate*	*95% credible region*
h^2_{y0}	0.0004	(0, 0.003)
h^2_{y1}	0.0006	(0, 0.006)
h^2_{y2}	0.0003	(0, 0.004)
$\sigma^2_{\epsilon(0)}$	1148.0	(953.5, 1373.0)
$\sigma^2_{\epsilon(1)}$	95.54	(79.83, 114.5)
$\sigma^2_{\epsilon(2)}$	0.080	(0.067, 0.096)
μ_0	93.07	(−23.22, 172.3)
μ_1	52.07	(42.37, 61.19)
μ_2	2.723	(2.364, 3.082)
block (0,1)	72.93	(–5.187, 189.8)
block (0,2)	73.48	(–4.655, 190.0)
block (0,3)	81.36	(–0.659, 200.1)
block (0,4)	76.15	(–3.133, 195.3)
block (1,1)	−0.814	(–7.545, 7.120)
block (1,2)	−3.933	(–11.07, 3.989)
block (1,3)	0.355	(–6.416, 8.406)
block (1,4)	6.842	(0.301, 15.94)
block (2,1)	−0.165	(–0.402, 0.077)
block (2,2)	−0.043	(–0.277, 0.202)
block (2,3)	0.110	(–0.120, 0.359)
block (2,4)	0.095	(–0.138, 0.362)
$\sigma^2_{\text{block}[0]}$	21990.0	(0.022, 131600.0)
$\sigma^2_{\text{block}[1]}$	57.39	(4.600, 291.6)
$\sigma^2_{\text{block}[2]}$	0.066	(0.007, 0.317)

The posterior estimates (mean and 95% credible region) of heritabilities, residual variances, the baselines for the three latent traits (intercept $y0$, slope $y1$ and quadratic term $y2$). Additionally, the posterior estimates for the block coefficients and their corresponding block variances are shown for the three latent traits.

Discussion

A conceptual description of the new method for mapping functional QTLs was presented in this paper. Because of its conceptual nature, it is worth emphasising that practical and scalable implementations of the method are out of the scope of this very first paper. The method is based on mapping QTLs which influence the curve parameters (that is, latent traits) describing functional curve of time-dependent phenotypic measurements (cf. Gee *et al*., 2003; Heuven and Janss, 2010). Unlike the others, we use a multitrait multiple-QTL model, which is essentially a single Bayesian-hierarchical model allowing for information flow and incorporation of uncertainties between different levels of the hierarchy. Also, the multitrait analysis is known to improve the accuracy and power of QTL detection (Jiang and Zeng, 1995). Note that Gee *et al*. (2003) and Heuven and Janss (2010) used two-stage approaches and performed QTL analysis for each parameter of the curve separately. One possible application field of the hierarchical model presented here is to map regulatory loci controlling time-dependent changes of gene expressions (eQTL, transcript abundances) or protein expressions over time (pQTL) (Reis *et al*., 2001; Foss *et al*., 2007; Ge *et al*., 2010). In such applications, eQTLs can influence the curve parameters, which again determine the individual's expression curve over time. However, in these situations, the current polynomial curve function may not be flexible enough to describe the required nonlinear shape of the expression profile (see Luan and Li, 2004; Qu and Xu, 2006). Of course, this unsuitability can be somewhat handled by giving high value to σ^2_ϵ but at the same time,

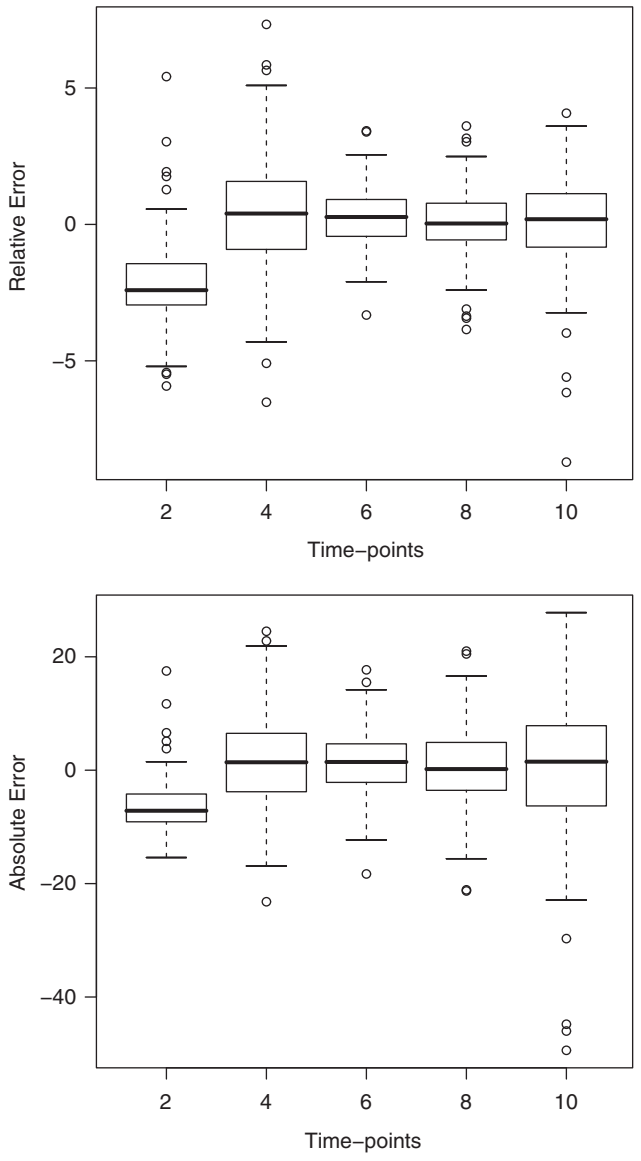

Figure 2 The prediction accuracy of the phenotypes in the Scots pine data. Boxplots of relative errors (top) and absolute errors (bottom) for predicted phenotypes at five time points, which had increased missingness. Errors are shown on the y-axis and the time-points on the x-axis. The absolute error is calculated as difference between predicted (posterior mean) and true phenotype and relative error is obtained as 100 times absolute error divided by the absolute value of the true phenotype.

it has a negative influence on statistical power to find QTLs. Thus, one may need to replace the model (1) with a more flexible function, for example, one which is possibly first estimated based on available set of featured genes from the biological process in question (Luan and Li, 2004).

There is recent interest in using recursive relationships or feedback effects in multitrait quantitative genetic models (Gianola and Sorensen, 2004; Wu *et al*., 2010). Thus, we shortly comment on differences between such recursive models and our autoregressive formulation of multitrait QTL model. First, our model assumes residual independence while these recursive models assume residual dependence. Second, we have effects of trait 1 to trait 2 and trait 3 but not the other way round while recursive models may include all the effects or cyclic dependencies between the traits. Although it is possible to include more complicated between-trait interdependencies into the model, it is not well justified in our setting. One should remember that our decision to include autoregressive coefficients to the QTL models (2–4) was made to improve computational efficiency only.

The multitrait multiple-QTL part constitutes one level of our larger hierarchical multilevel model. It represents a new and efficient way to handle residual dependencies between quantitative traits. How this part of the model performs in the presence of a larger number of traits (for example, higher order polynomials) deserves to be more carefully studied in the future. Moreover, it is still an open question here what kind of modifications are needed for models (2–4) when the trait values are observed quantities rather than model parameters.

As seen in examples, the presented hierarchical model can be used to predict unobserved phenotypic measurements at arbitrary time points from posterior predictive distributions. The curve parameters (and underlying QTL model parameters) estimated based on observed measurements from all individuals provide information for predicting a single time-point of an individual. As collecting phenotype data is expensive, one application of the model may be to use this feature in data collection. Thus, one can systematically reduce the number of individuals collected at some of the time points by randomly selecting measured individuals at each point. To maintain accuracy in the curve parameters, as illustrated in our examples, one can collect systematically more complete data sample at every other measurement point. However, it is important to keep in mind here that the hierarchical model does not represent informative missing data model similarly as in Sillanpää and Noykova (2008), because missing values occur here at the highest level of the hierarchy in the model. This means that only the observed part of phenotypes over time points have influence on the posterior distributions of the model parameters. Therefore, the degree of missingness at any time point should not fluctuate too much from other time points.

In our analysis of simulated data replicates, we found that our method had problems to find QTLs for the intercept while analysis for the other two latent traits worked much better. Weak signals were also found in real data analysis while our method worked clearly better for the simulated QTLMAS data set, where the sample size was relatively large. However, our suggested position of the second major QTL in chromosome 1 could have been more accurate with larger sample. These results may indicate the importance of having large sample size in functional QTL studies. On the other hand, our method was able to provide accurate phenotype predictions with small data.

The method was implemented using WinBUGS software, which allows MCMC estimation of the hierarchical model parameters without requiring derivation of the details of the sampling algorithm such as fully conditional posterior distributions. When the size of the marker sets increases and/or the WinBUGS implementation becomes too slow for the practical purposes, one can

proceed by (i) implemenenting one's own MCMC sampler using a convenient programming language or (ii) perform pre-selection of the marker set to reduce the model dimensionality. For the variable selection part of our hierarchical model, we recommend relying on some existing sampling algorithms and their full conditionals, which may guarantee the sufficient mixing properties of the sampler. For example, see Banerjee *et al*., 2008 for implementational details of suitable MCMC sampling algorithm in this respect. The sampling steps for individual-specific functional curve parameters can then be included as additional steps into the sampling scheme of Banerjee *et al*., 2008. The full conditional distributions of functional curve parameters have analytical forms because of conjugacy: both the likelihood and the prior are densities of the normal distribution (details not shown). The pre-selection of the markers can be carried out for example, by first estimating the curve parameters and then eliminating the markers that are weakly correlated with each curve parameter with a single-marker test (see Cho *et al*., 2010). Alternatively, one can use pre-selected set of haplotype-tagging markers in the analysis (see Lin and Altman, 2004).

In *Pinaceae*, age-to-age correlations and narrow-sense heritability for height have generally been reported to be low to moderate with an increasing tendency with age (Lambeth, 1980; Costa and Durel, 1996; Jansson *et al*., 2003, 2005, but see Gwaze, 2009). Low correlation across ages has also been observed for QTL identification (Plomion *et al*., 1996; Verhaegen *et al*., 1997; Kaya *et al*., 1999); this is partly due to different set of genes expressed at each life stage, although environmental variance is expected to be large, especially at early growth stages. Although, a time-point QTL analysis of Lerceteau *et al*. (2001) reported consistency in the number and location of QTLs for height across four years in Scots pine, trees were still at the juvenile stage and QTL expression at mature stages was not verified. A study of Lerceteau *et al*. (2001) was based on 94 individuals and a total of 152 dominant markers (59 maternal and 93 paternal), but their marker set was different from our marker set here which makes the comparison difficult. However, we also detected three QTLs in our study when analysing functional growth data for 11 years in the same Scots pine population. Although a similar number of QTLs was found in both studies, QTLs may not be equivalent as we used a functional QTL-mapping approach. A functional QTL mapping is an alternative approach that focuses on the developmental features of the dynamic trait (for example, growth curve) overcoming the problem of age-specific QTL expression. Many studies in conifers have been devoted to analyse growth trajectories (see Balocchi *et al*., 1993; Magnussen and Kremer, 1993; Danjon, 1994; Gwaze *et al*., 2002; Wang *et al*., 2009). However, to our knowledge, no QTL analysis in conifers has been trying to investigate functional traits, the only published works being in Populus (Wu *et al*., 2003; Ma *et al*., 2004). Our analysis revealed three QTLs for growth parameters such as slope (speed of growth) or quadratic term (curvature or timing of growth cessation), which are essential for the genetic improvement of forest trees and can only be assessed by means of dynamic trait analysis. Growth curve parameter estimation has critical advantages such as the fit of the data to a biologically meaningful mathematical model, which furthermore helps to correct for data irregularities due to human errors or environmental effects. Furthermore, dynamic trait analysis could also be useful to predict growth at ages where measurements are missing. Growth trajectory parameters can be shifted as a response to selection. Breeding on growth curves are used in animal breeding (Tholon and de Queiroz, 2009; Haraldsen *et al*., 2009) and the same results could be expected when used in forest tree breeding.

In our hierarchical model, additive genetic variation (of QTLs) influences the curve parameters, which in turn control the shape of the polynomials over time. In this context, we illustrated estimation of additive genetic variances and heritabilites for these curve parameters and genetic covariances between them. Unlike the common practice (Gwaze *et al*., 2002; Kulathinal *et al*., 2008; Wang *et al*., 2009), our analysis does not provide time-specific heritability or covariance estimates at all. However, it may be more meaningful from a breeding point of view to actually inspect the genetics and estimate the genomic breeding values underlying the curve characteristics (which control the dynamic behaviour of the trait), rather than inspecting genetics at different time-points. For example, the slope will be easy to interpret from a biological point of view as 'speed of growth'. Note that the additive genetic variance was estimated as the variance of the genomic breeding values which provided a marker-based estimate for heritability (cf. Meuwissen *et al*., 2001; Xu, 2003; Pikkuhookana and Sillanpää, 2009; Sillanpää, 2011). Generally, these heritability estimates (when the same growth function was used in simulation and analysis) were underestimated due to presumably small sample size, but the accuracy of the predicted phenotypes were, especially high, and they both will motivate future studies.

The model specification codes (written in WinBUGS) used in this article and instructions to use them are freely available for research purposes at URL http://www.rni.helsinki.fi/~mjs/.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

We are grateful to Crispin M Mutshinda for useful discussions and valuable comments on the manuscript and three anonymous reviewers for their constructive comments on the manuscript. This work was supported by a research grant from the Academy of Finland, University of Helsinki's Research Funds, Research of Forest Genetics and Breeding, Kempe Foundation and by the Research School of Forest Genetics at the Swedish University of Agriculture, SLU.

References

Balocchi CE, Bridgwater FE, Zobel BJ, Jahromi S (1993). Age trends in genetic-parameters for tree height in a nonselected population of loblolly-pine. *For Sci* **39**: 231–251.

Banerjee S, Yandell BS, Yi N (2008). Bayesian quantitative trait loci mapping for multiple traits. *Genetics* **179**: 2275–2289.

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

Cho S, Kim K, Kim YJ, Lee J-K, Cho YS, Lee J-Y *et al*. (2010). Joint identification of multiple genetic variants via elastic-net variable selection in a genome-wide association analysis. *Ann Hum Genet* **74**: 416–428.

Conner PJ, Brown SK, Weeden NF (1998). Molecular-marker analysis of quantitative traits for growth and development in juvenile apple trees. *Theor Appl Genet* **96**: 1027–1035.

Costa P, Durel CE (1996). Time trends in genetic control over height and diameter in maritime pine. *Can J For Res* **26**: 1209–1217.

Coster A, Bastiaansen JWM, Calus MPL, Maliepaard C, Bink MCAM (2010). QTLMAS 2009: simulated dataset. *BMC Proc* **4**(Suppl 1): 53.

Danjon F (1994). Heritabilities and genetic correlations for estimated growth curve parameters in maritime pine. *Theor Appl Genet* **89**: 911–921.

Druet T, Georges M (2010). A hidden Markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. *Genetics* **184**: 789–798.

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

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

Ge H, Wei M, Fabrizio P, Hu J, Cheng C, Longo VD *et al*. (2010). Comparative analyses of time-course gene-expression profiles of the long-lived sch9Δ mutant. *Nucleic Acids Res* **38**: 143–158.

Gee C, Morrison JL, Thomas DC, Gauderman WJ (2003). Segregation and linkage analysis for longitudinal measurements of a quantitative trait. *BMC Genetics* **4**(Suppl 1): S21.

Gianola D, Sorensen D (2004). Quantitative genetic models for describing simultaneous and recursive relationships between phenotypes. *Genetics* **167**: 1407–1424.

Gwaze D (2009). Optimum selection age for height in shortleaf pine. *New Forests* **37**: 9–16.

Gwaze DP, Bridgwater FE, Williams CG (2002). Genetic analysis of growth curves for a woody perennial species, *Pinus taeda L. Theor Appl Genet* **105**: 526–531.

Haraldsen M, Odegard J, Olsen D, Vangen O, Ranberg IMA, Meuwissen THE (2009). Prediction of genetic growth curves in pigs. *Animal* **3**: 475–481.

Heffner EL, Sorrels ME, Jannink JL (2009). Genomic selection for crop improvement. *Crop Sci* **49**: 1–12.

Heuven HCM, Janss LLG (2010). Bayesian multi-QTL mapping for growth curve parameters. *BMC Proc* **4**(Suppl 1): S12.

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

Jansson G, Jonsson A, Eriksson G (2005). Use of trait combinations for evaluating juvenile-mature relationships in *Picea abies* (L). *Tree Genet Genomes* **1**: 21–29.

Jansson G, Li B, Hannrup B (2003). Time trends in genetic parameters for height and optimal age for aprental selection in Scots pine. *For Sci* **49**: 696–705.

Jeffreys H (1961). *Theory of Probability* 3rd edn. Claredon Press Oxford: UK.

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

Kass RE, Raftery AE (1995). Bayes factors. *J Am Stat Assoc* **90**: 773–795.

Kaya Z, Sewell MM, Neale DB (1999). Identification of quantitative trait loci influencing annual height- and diameter-increment growth in loblolly pine (*Pinus teada L. Theor Appl Genet* **98**: 586–592.

Kulathinal S, Gasbarra D, Kinra S, Ebrahim S, Sillanpää MJ (2008). Estimation of additive genetic and environmental sources of quantitative trait variation using data on married couples and their siblings. *Genet Res* **90**: 269–279.

Lambeth CC (1980). Juvenile-mature correlations in *Pinaceae* and implications for early selection. *For Sci* **26**: 571–580.

Lee SH, van der Werf JHJ, Hayes BJ, Goddard ME, Visscher PM (2008). Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *PloS Genet* **4**: e1000231.

Lerceteau E, Szmidt AE, Andersson B (2001). Detection of quantitative trait loci in *Pinus sylvestris L.* across years. *Euphytica* **121**: 117–122.

Lin Z, Altman RB (2004). Finding haplotype tagging SNPs by use of principal component analysis. *Am J Hum Genet* **75**: 850–861.

Lorenzana RE, Bernando R (2009). Accuracy of genotypic value prediction for marker-based selection in biparental plant populations. *Theor Appl Genet* **120**: 151–161.

Luan Y, Li H (2004). Model-based method for identifying periodically expressed genes based on time course microarray gene expression data. *Bioinformatics* **20**: 332–339.

Lund M, Sorensen P, Madsen P, Jaffrézic F (2008). Detection and modelling of time-dependent QTL in animal populations. *Genet Sel Evol* **40**: 177–194.

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

Ma CX, Casella G, Wu RL (2002). Functional mapping of quantitative trait loci underlying the character process: a theoretical framework. *Genetics* **161**: 1751–1762.

Ma CX, Lin M, Littell RC, Yin T, Wu RL (2004). A likelihood approach for mapping growth trajectories using dominant markers in a phase-unknown full-sib family. *Theor Appl Genet* **108**: 699–705.

Macgregor S, Knott SA, White I, Visscher PM (2005). Quantitative trait locus analysis of longitudinal trait data in complex pedigrees. *Genetics* **171**: 1365–1376.

Magnussen S, Kremer A (1993). Selection for an optimum tree growth curve. *Silvae Genet* **42**: 322–335.

Marchini J, Howie B (2010). Genotype imputation for genome-wide association studies. *Nat Revs Genet* **11**: 499–511.

Martinez V, Thorgaard G, Robison B, Sillanpää MJ (2005). An application of Bayesian QTL mapping to early development in double haploid lines of rainbow trout including environmental effects. *Genet Res* **86**: 209–221.

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

Min L, Yang R, Wang X, Wang B (2011). Bayesian analysis of genetic architecture of dynamic traits. *Heredity* **106**: 124–133.

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

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

Piyasatian N, Fernando RL, Dekkers JCM (2007). Genomic selection for marker-assisted improvement in line crosses. *Theor Appl Genet* **115**: 665–674.

Pletcher SD, Geyer C (1999). The genetic analysis of age-dependent traits: modeling the character process. *Genetics* **153**: 825–835.

Plomion C, Durel CE, O'Malley DM (1996). Genetic dissection of height in maritime pine seedlings raised under accelerated growth conditions. *Theor Appl Genet* **93**: 849–858.

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

Reis BY, Butte AS, Kohane IS (2001). Extracting knowledge from dynamics in gene expression. *J Biomed Inform* **34**: 15–27.

Robert C, Casella G (2004). *Monte Carlo Statistical Methods* 2nd edn. Springer-Verlag: New York.

Servin B, Stephens M (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. *PloS Genet* **3**: e114.

Sillanpää MJ (2011). On statistical methods for estimating heritability in wild populations. *Mol Ecol* **20**: 1324–1332.

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

Sillanpää MJ, Arjas E (1999). Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. *Genetics* **151**: 1605–1619.

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

Spiegelhalter D, Thomas A, Best N, Lunn D (2005). *WinBugs User Manual*, Version 2.10 MRC Biostatistics Unit, Institute of Public Health: Cambridge, UK.

Tholon P, de Queiroz SA (2009). Mathematic models applied to describe growth curves in poultry applied to animal breeding. *Ciencia Rural* **39**: 2261–2269.

Verhaegen D, Plomion C, Gion JM, Poitel M, Costa P, Kremer A (1997). Quantitative trait dissection analysis in Eucalyptus using RAPD markers: 1. Detection of QTL in interspecific hybrid progeny stability of QTL expression across different ages. *Theor Appl Genet* **95**: 597–608.

Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M *et al*. (1995). AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res* **23**: 4407–4414.

Wang C, Andersson B, Waldmann P (2009). Genetic analysis of longitudinal height data using random regression. *Can J For Res* **39**: 1939–1948.

West GB, Brown JH, Enqvist BJ (2001). A general model for ontogenetic growth. *Nature* **413**: 628–631.

Wu RL, Ma CX, Chang M, Littell RC, Wu SS, Yin TM *et al*. (2002). A logistic mixture model for characterizing genetic determinants causing differentiation in growth trajectories. *Genet Res* **79**: 235–245.

Wu RL, Ma CX, Min L, Casella G (2004). A general framework for analyzing the genetic architecture of developmental characteristics. *Genetics* **166**: 1541–1551.

Wu RL, Ma CX, Yhang M, Chang M, Littell RC, Santra U *et al*. (2003). Quantitative trait loci for growth trajectories in *Populus*. *Genet Res* **81**: 51–64.

Wu RL, Lin M (2006). Functional mapping - how to map and study the genetic architecture of dynamic complex traits. *Nat Revs Genet* **7**: 229–237.

Wu X-L, Heringstad B, Gianola D (2010). Bayesian structural equation models for inferring relationships between phenotypes: a review of methodology, identifiability, and applications. *J Anim Breed Genet* **127**: 3–15.

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

Yang R, Tian Q, Xu S (2006). Mapping quantitative trait loci for longitudinal traits in line crosses. *Genetics* **173**: 2339–2356.

Yang R, Xu S (2007). Bayesian shrinkage analysis of quantitative trait loci for dynamic traits. *Genetics* **176**: 1169–1185.

Yi N, Shriner D, Banerjee S, Mehta T, Pomp D, Yandell BS (2007). An efficient Bayes model selection approach for interacting quantitative trait loci models with many effects. *Genetics* **176**: 1865–1877.

Yi N, Xu S (2008). Bayesian LASSO for quantitative trait loci mapping. *Genetics* **179**: 1045–1055.