

SIMULTANEOUS SNP IDENTIFICATION IN ASSOCIATION STUDIES WITH MISSING DATA

BY ZHEN LI, VIKNESWARAN GOPAL, XIAOBO LI,
JOHN M. DAVIS AND GEORGE CASELLA

*State Street Corporation, University of Florida, University of Florida,
University of Florida and University of Florida*

Association testing aims to discover the underlying relationship between genotypes (usually Single Nucleotide Polymorphisms, or SNPs) and phenotypes (attributes, or traits). The typically large data sets used in association testing often contain missing values. Standard statistical methods either impute the missing values using relatively simple assumptions, or delete them, or both, which can generate biased results. Here we describe the Bayesian hierarchical model BAMD (Bayesian Association with Missing Data). BAMD is a Gibbs sampler, in which missing values are multiply imputed based upon all of the available information in the data set. We estimate the parameters and prove that updating one SNP at each iteration preserves the ergodic property of the Markov chain, and at the same time improves computational speed. We also implement a model selection option in BAMD, which enables potential detection of SNP interactions. Simulations show that unbiased estimates of SNP effects are recovered with missing genotype data. Also, we validate associations between SNPs and a carbon isotope discrimination phenotype that were previously reported using a family based method, and discover an additional SNP associated with the trait. BAMD is available as an R-package from <http://cran.r-project.org/package=BAMD>.

1. Introduction. This work was motivated from a study of the genomics of loblolly pine, an economically and ecologically important tree species in the United States. The native range of loblolly pine extends from Maryland, south to Florida, and west to Texas. Its annual harvest value is approximately 19 billion dollars [McKeever and Howard (1996)]. The pine species in the southern states produces 58% of the timber in the US and 15.8% of the world's timber [Wear and Greis (2002)]. We are interested in discovering

Received December 2010; revised September 2011.

Key words and phrases. Hierarchical models, Bayes models, Gibbs sampling, genome-wide association.

This is an electronic reprint of the original article published by the Institute of Mathematical Statistics in *The Annals of Applied Statistics*, 2012, Vol. 6, No. 2, 432–456. This reprint differs from the original in pagination and typographic detail.

the relationship between phenotypic traits and genes underlying complex traits in loblolly pine, so we can understand their evolution and apply that knowledge to genetic improvement. We are especially interested in SNPs associated with disease resistance and response to water deficit.

Large genomic data sets typically contain missing data. Missing data create imbalance and complicate calculations required for statistical analyses. There are various approaches to dealing with missing data. Eliminating cases is one approach, but undesirable in large data sets where most or all cases have missing data. Imputation is more commonly used [Huisman (2000), Dai et al. (2006)]. Single imputation using haplotype data [Marchini et al. (2007), Su et al. (2008), Sun and Kardia (2008), Szatkiewicz et al. (2008)], either implicitly or explicitly, relies on linkage disequilibrium among markers, or information that can be extracted from other data sets [Stephens, Smith and Donnelly (2001), Scheet and Stephens (2006), Servin and Stephens (2007)]. However, there is no reference genome sequence for loblolly pine, so it is not possible to impute missing SNPs from flanking SNPs.

It is well established that single imputation approaches, while fast, can give biased parameter estimates [Greenland and Finkle (1995); see also van der Heijden et al. (2006)]. The best approach is to average over the missing data using the formal missing data distribution, rather than to impute a single value based on a possibly ad hoc scheme. This is appealing because it addresses uncertainty and variability in the missing data [Little and Rubin (2002), Dai et al. (2006)], particularly in species or genomic regions where LD decays rapidly and thus adjacent SNPs are not necessarily correlated [Flint-Garcia, Thornsberry and Buckler (2003), Neale and Ingvarsson (2008)]. However, multiple imputation is so computationally intensive that, prior to the present work, it has not been feasible for larger genomic data sets.

Several approaches have been developed recently to enable association testing. Association testing identifies relationships between polymorphisms in DNA sequence (most commonly Single Nucleotide Polymorphisms, or SNPs) and phenotypes, as a strategy to identify the genes that control traits [Flint-Garcia, Thornsberry and Buckler (2003), Hirschhorn and Daly (2005), Balding (2006)]. For family-based analysis, Chen and Abecasis (2007) used an identity by descent parameter to measure correlation among SNPs, and a kinship coefficient to model the correlation among siblings to develop the Quantitative Transmission Disequilibrium Test (QTDT). Other approaches allow association testing in populations with recent mating or historical (unrecorded) mating, or combinations. TASSEL fits a mixed model to detect associations while taking into account both “coarse-scale” relatedness based on population structure [Pritchard, Stephens and Donnelly (2000)] and “fine-scale” relatedness based on a matrix of kinship coefficients [Yu et al. (2006)]. Our approach for family based analyses accomplishes the same goal by employing the numerator relationship matrix [Henderson (1976); see

also Quaas (1976)], which avoids complications arising from nonpositive definite matrices derived from complex interrelationships. A desirable feature of any association testing approach is simultaneous solution of multiple SNP effects to prevent upward bias in parameter estimates, and to appropriately model the underlying biological system in which many SNPs act in concert to condition the phenotype. Such an approach is developed in Wilson et al. (2010), who introduce Multilevel Inference for SNP Association (MISA), using imputation from fastPHASE [Stephens, Smith and Donnelly (2001), Servin and Stephens (2007)] and a Bayes-driven stochastic search method to find good models.

In this paper we introduce BAMD (Bayesian Association with Missing Data) and show that computation time required for formal multiple imputation can be reduced without sacrificing accuracy, establishing the feasibility of using BAMD on genomic data sets. Our approach is to use all available data in imputation of missing SNPs. This approach is motivated statistically; we use all of the available information to estimate SNP effects on phenotypes, across all possible values for missing SNPs. Prior knowledge such as pedigree structure may be used as constraints. Simulations show that BAMD detects SNP effects efficiently.

On the loblolly pine genomic data, we used a series of *tag SNPs* [González-Martínez et al. (2008)]. Tag SNPs are markers that are relatively evenly dispersed throughout the genome, and are used to survey chromosomal segments for genes that underly phenotypes. One assessment of the performance of BAMD was to use it on this same population, genotype, and phenotype data, in which it had been found that three of the tag SNPs were significantly associated with carbon isotope discrimination [a measure of water use efficiency, González-Martínez et al. (2008)]. BAMD detected a fifth tag SNP in addition to the other four tag SNPs that were detected using QTDT in that previous work.

An additional feature of BAMD is the variable selector. The variable selector searches model space for the most parsimonious set of SNPs that explain the phenotype. This feature is designed for unsupervised discovery of interactions among SNPs, and should find application in situations where epistatic interactions are important determinants of phenotype.

The remainder of the paper is organized as follows. In Section 2 we describe the model and the estimation of parameters, and Section 3 describes the variable selector, including the use of Bayes factors, the stochastic search, and computational strategies. In Section 4.1, we investigate the amount of missing data that BAMD can handle through a simulation. Section 4.2 compares our procedure to BIMBAM, a popular genomics program that does both imputation and variable selection. Section 5 analyzes the loblolly pine data, where we discover a previously undiscovered SNP. Section 6 contains a concluding discussion. Computational implementation is described in the

[Appendix](#), and the accompanying theorems and proofs can be found in the online Supplemental Information [Li et al. (2011)].

2. Model. Our method can be viewed as a two-stage procedure. The first stage involves identifying individual SNPs that have significant effects on the phenotype, with all SNPs in the model. The second stage searches for the best subset of SNPs, from those picked out in the first stage. First we describe the model.

2.1. *Conceptual framework for BAMD.* The response is assumed to be continuous, following a normal distribution. The data set has fully observed family covariates for all the observations. Missing values are imputed only among SNPs, although the method can be modified to impute missing values for phenotypes as well. We focus on testing the relationship between the response and the SNPs. We assume only additive effects among SNPs, although the method can be adapted to quantifying additive and dominance effects of SNPs.

We begin with the linear mixed model

$$(1) \quad \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

where $\mathbf{Y}_{n \times 1}$ is the phenotypic trait, $\mathbf{X}_{n \times p}$ is the design matrix for family covariates, $\boldsymbol{\beta}_{p \times 1}$ are the coefficients for the family effect, $\mathbf{Z}_{n \times s}$ is the design matrix for SNPs (genotypes), $\boldsymbol{\gamma}_{s \times 1}$ are the coefficients of the additive effect for SNPs, and $\boldsymbol{\varepsilon}_{n \times 1} \sim N(0, \sigma^2 \mathbf{R})$. The matrix \mathbf{R} is the numerator relationship matrix, describing the degree of kinship between different individuals. (Details on the calculation of \mathbf{R} are given in Appendix A.)

For our application here (carbon isotope data), we have $n = 1000$ and $s = 450$. In another application of BAMD [Quesada et al. (2010)], they used $n = 450$ and $s = 400$. In both cases the number of covariates, p , was less than 6. With a fully Bayesian implementation, it is possible to adapt BAMD to the $p \gg n$ case.

Each row of the matrix \mathbf{Z} , $\mathbf{Z}_i, i = 1, \dots, n$, corresponds to the SNP genotype information of one individual, which can be homozygous for either of the two nucleotides $(-1, 1)$ or heterozygous (0) . Some of this information may be missing, and we write $\mathbf{Z}_i = (\mathbf{Z}_i^{\text{obs}}, \mathbf{Z}_i^{\text{miss}})$, where $\mathbf{Z}_i^{\text{obs}}$ are the observed genotypes for the i th individual, and $\mathbf{Z}_i^{\text{miss}}$ are the missing genotypes. Note two aspects of this framework:

(1) The values of $\mathbf{Z}_i^{\text{miss}}$ are not observed. Thus, if $*$ denotes one missing SNP, a possible \mathbf{Z}_i is $\mathbf{Z}_i = (1, *, 0, 0, *, *, 1)$.

(2) Individuals are likely to have missing data at different SNP loci. So for 2 different individuals, we might have

$$\mathbf{Z}_i = (1, *, 0, 0, *, *, 1) \quad \text{and} \quad \mathbf{Z}_{i'} = (*, *, 1, 0, 0, 1, 1).$$

For a Bayesian model, we want to put prior distributions on the parameters. We put a noninformative uniform prior for $\boldsymbol{\beta}$, which essentially leads us

to least squares estimation. For $\boldsymbol{\gamma}$, we use the normal prior $\boldsymbol{\gamma} \sim N(\mathbf{0}, \sigma^2 \phi^2 \mathbf{I}_s)$. Here ϕ^2 is a scale parameter for the variance and σ^2 is the variance parameter. For σ^2 and ϕ^2 , we use inverted Gamma priors: $\sigma^2 \sim \text{IG}(a, b)$ and $\phi^2 \sim \text{IG}(c, d)$, where IG stands for the inverted Gamma distribution, and a, b, c , and d are constants used in the priors. For specified a, b, c, d , the resulting posterior distribution is proper [see Hobert and Casella (1996)].

We consider the case of tag SNP markers in the loblolly pine genome with no significant linkage disequilibrium between them [González-Martínez et al. (2006)]. Therefore, noninformative priors are used for the missing SNPs, meaning that missing data have equal probability of any allelic state. As information increases due to higher marker density, or parental information, or allele frequency in the population, missing data imputation could be constrained accordingly.

We assume that missing SNPs in the data set are *Missing at Random (MAR)*. In particular, let the value of the random variable T denote whether Z is observed, with $T = 1$ if the value is observed and $T = 0$ if it is missing. If ξ is the parameter of the missing mechanism, then, under the model (1), the MAR assumption results in

$$P(T|Y, Z^{\text{obs}}, Z^{\text{miss}}, \xi) = P(T|Y, Z^{\text{obs}}, \xi).$$

So the distribution of the missing SNP could depend on the observed SNPs, and the observed phenotypes. Of course, this does not require such a dependence, it only allows for it.

Other assumptions about missing data mechanisms are less common than MAR. The strongest assumption, and the most difficult to justify, is *Missing Completely at Random (MCAR)*. Under this assumption, the missing data distribution is independent of all observed data, the complete cases can be regarded as sub-samples from the population, and statistical inference with regard to the complete cases is totally valid. Under the model (1), the MCAR assumption can be expressed as

$$P(T|Y, Z^{\text{obs}}, Z^{\text{miss}}, \xi) = P(T|\xi).$$

MCAR is regarded as unrealistic and, in most cases, it is not satisfied. It is typically not used to model missing data, and we do not use it here.

Conditional on the MAR assumption, we impute the missing SNPs based on the correlation between SNPs within individuals and between individuals, and use the phenotypic trait information to improve the power of imputation.

In this model, the covariance matrix, \mathbf{R} , models the covariance between individuals within the same family, and covariance between individuals across families. Phenotypic traits of related individuals are alike because they share some proportion of SNPs, and genotypes of relatives are similar because they share the same alleles passed on from parents. Various methods can be used to calculate the relationship matrix, such as using a co-ancestry matrix,

a kinship matrix, etc. The basic idea is to calculate the probability that 2 individuals share SNPs that are identical by descent. Some methods use pairwise calculations and thus do not guarantee a positive definite relationship matrix, which is unsatisfactory when the relationship matrix is used as covariance matrix. We use the recursive calculation method of Henderson (1976), which gives a numerator relationship matrix that quantifies the probability of sharing a SNP from the same ancestry, based on known family pedigree and parent pedigree in the population. So by calculating this relationship matrix we obtain a probability of 0.5 for the case that two siblings are within the same control-pollinated family and therefore share the same copy of a SNP, or a 0.25 probability if the two siblings only have one parent in common. For the complex pedigree that we analyze here, there are a total of 9 categories of relatedness.

2.2. Estimation of parameters. The model (1) along with the prior specification allows the use of a Gibbs sampler to estimate parameters. We can iteratively sample from the the full conditionals, given by

$$\begin{aligned}
\beta &\sim N((X'R^{-1}X)^{-1}X'R^{-1}(Y - \mathbf{Z}\gamma), \sigma^2(X'R^{-1}X)^{-1}), \\
\gamma &\sim N\left(\left(\mathbf{Z}'R^{-1}\mathbf{Z} + \frac{I}{\phi^2}\right)^{-1}\mathbf{Z}'R^{-1}(Y - X\beta), \sigma^2\left(\mathbf{Z}'R^{-1}\mathbf{Z} + \frac{I}{\phi^2}\right)^{-1}\right), \\
(2) \quad \sigma^2 &\sim \frac{1}{(\sigma^2)^{n/2+s/2+a+1}} \\
&\quad \times \exp\left(-\frac{(Y - X\beta - \mathbf{Z}\gamma)'R^{-1}(Y - X\beta - \mathbf{Z}\gamma) + |\gamma|^2/\phi^2 + 2b}{2\sigma^2}\right), \\
\phi^2 &\sim \frac{1}{(\phi^2)^{s/2+c+1}} \exp\left(-\frac{(|\gamma|^2/\sigma^2 + 2d)}{2\phi^2}\right).
\end{aligned}$$

The SNPs are contained in the \mathbf{Z} matrix, which includes both the observed SNPs and missing SNPs, and we use the Gibbs sampler to impute the missing SNPs. The Gibbs sampler for the missing data simulates the samples of Z_i^m according to the distribution of each missing SNP conditional on the rest of observed SNPs and sampled missing SNPs. For a particular SNP Z_{ij}^m , the j th missing SNP in the i th individual, the conditional distribution given the rest of the vector $Z_{i(-j)}^m$ and all other parameters in the model is

$$\begin{aligned}
(3) \quad &P(Z_{ij}^m = c | Z_{i(-j)}^m) \\
&= \frac{\exp(-(Y_i - X_i\beta - Z_i^o\gamma_i^o - Z_{i(-j)}^m\gamma_{i(-j)}^m - c\gamma_{ij}^m)^2/(2\sigma^2))}{\sum_{\ell=1}^3 \exp(-(Y_i - X_i\beta - Z_i^o\gamma_i^o - Z_{i(-j)}^m\gamma_{i(-j)}^m - c_\ell\gamma_{ij}^m)^2/(2\sigma^2))}.
\end{aligned}$$

The value c is the genotype currently being considered for that missing SNP, and c_ℓ represents any one of the possible genotypes for the SNP. Notice there

are only 3 terms in the denominator sum for each SNP and this is a key point why Gibbs sampling is computationally feasible for our situation with many SNPs and many observations. We also note that the EM algorithm, which provides an alternative method of parameter estimation, can require a prohibitive amount of computation. See Appendix B.

3. Variable selection. The Gibbs sampler will estimate the full set of parameters in model (1). However, it is often the case that a small set of SNPs will explain a sufficient proportion of the variability that might also be biologically meaningful. To this end, along with the Gibbs sampler, we run a second Markov chain that searches through the space of available models, looking for the one with the largest Bayes factor.

A model is specified by a vector δ of length s , whose entries are either 0 or 1. The γ vector of model (1) becomes $\gamma_\delta = \gamma \star \delta$, where “ \star ” denotes componentwise multiplication. The corresponding columns of \mathbf{Z} are deleted, giving \mathbf{Z}_δ , and the reduced model is

$$(4) \quad \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_\delta\boldsymbol{\gamma}_\delta + \boldsymbol{\varepsilon}.$$

Thus, the components of γ_i corresponding to $\delta_i = 0$ are excluded from the model. Correspondingly, let $\boldsymbol{\theta}$ denote the random vector consisting of all parameters in the full model, so $\boldsymbol{\theta} := (\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, \phi^2, \mathbf{Z})$ and, naturally, $\boldsymbol{\theta}_\delta := (\boldsymbol{\beta}, \boldsymbol{\gamma}_\delta, \sigma^2, \phi^2, \mathbf{Z}_\delta)$. Let m_δ, π_δ , and p_δ denote the marginal distribution of \mathbf{Y} , the prior distribution on $\boldsymbol{\theta}_\delta$, and the conditional distribution of \mathbf{Y} , respectively. We also write, if needed, $\boldsymbol{\theta} = (\boldsymbol{\theta}_\delta, \boldsymbol{\theta}_{\delta^c})$, the latter containing the remaining parameters not specified by δ . For the full model containing all parameters we omit the subscript.

3.1. *Searching with Bayes factors.* In order to compare models, we shall use the Bayes factor comparing each candidate model to the full model, given by

$$(5) \quad \text{BF}_\delta = \frac{m_\delta(\mathbf{Y})}{m(\mathbf{Y})} = \frac{\int \pi_\delta(\boldsymbol{\theta}_\delta) p_\delta(\mathbf{Y}|\boldsymbol{\theta}_\delta) d\boldsymbol{\theta}_\delta}{\int \pi(\boldsymbol{\theta}) p(\mathbf{Y}|\boldsymbol{\theta}) d\boldsymbol{\theta}},$$

where p denotes the full model. We now can compare models δ and δ' through their Bayes factors, as a larger Bayes factor corresponds to a model that explains more variability, when compared to the full model. These pairwise comparisons result in a consistent model selector [O’Hagan and Forster (2004)], and have an advantage over BIC, which is overly biased toward smaller models [Casella et al. (2009)].

We now set up a Metropolis–Hastings (MH) search that has target distribution proportional to the the Bayes factor, BF_δ . Given that we are at model δ , we choose a candidate δ' from a random walk (choose one component at random and switch $0 \rightarrow 1$ or $1 \rightarrow 0$) with probability a and, with probability $1 - a$, we do an independent jump. This is a symmetric candidate, and δ' is accepted with probability $\min\{1, \text{BF}_{\delta'}/\text{BF}_\delta\}$.

3.2. *Estimating the Bayes factor.* Calculating the Bayes factor in (5) requires knowing the \mathbf{Z} matrix, which is not the case with missing data. Thus, to calculate the Bayes factor, we need to use the imputed \mathbf{Z} matrix from the Gibbs sampler. Thus, we run two Markov chains simultaneously:

(1) A Gibbs sampler on the full model, to impute the missing data in \mathbf{Z} and estimate all parameters.

(2) A Metropolis–Hastings algorithm on δ , in model space, to find the best model. This algorithm uses an estimated Bayes factor based on the current values in the Gibbs chain.

The aim is to search for δ^* such that $\delta^* = \arg \max_{\delta} \text{BF}_{\delta}$, but since we are not able to compute BF_{δ} exactly for any given δ , we estimate it using samples from the Gibbs sampler, which yields a strongly consistent estimator. We then use the estimated Bayes factor as the target in a stochastic search driven by a Metropolis–Hastings algorithm.

A typical method of estimating a quantity such as (5) would be to use bridge sampling [Meng and Wong (1996)]. However, since the numerator and denominator have different dimensions (but the numerator model is always nested in the denominator model), ordinary bridge sampling will not work. A variation [Chen and Shao (1997)] which accounts for this introduces a weight function to handle the dimension difference. We summarize this strategy in the following proposition.

PROPOSITION 1. *Referring to (5), let $g(\boldsymbol{\theta})$ be such that $\int p(\mathbf{Y}|\boldsymbol{\theta})g(\boldsymbol{\theta}) d\boldsymbol{\theta}_{\delta^c} = p_{\delta}(\mathbf{Y}|\boldsymbol{\theta}_{\delta})$. Then if expectation is taken with respect to the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{Y})$,*

$$\mathbb{E} \left[\frac{\pi_{\delta}(\boldsymbol{\theta}_{\delta})g(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})} \right] = \text{BF}_{\delta}.$$

One particular g function is defined as follows. Let $P_{\delta^c} := \mathbf{Z}_{\delta^c}'(\mathbf{Z}_{\delta^c}'\mathbf{Z}_{\delta^c})^{-1}\mathbf{Z}_{\delta^c}'$, $\mathbf{C}_{\delta} := (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}_{\delta}\boldsymbol{\gamma}_{\delta})$, and

$$(6) \quad g(\boldsymbol{\theta}) = (2\pi\sigma^2)^{-d^c/2} |\mathbf{Z}_{\delta^c}'\mathbf{Z}_{\delta^c}|^{1/2} \times \exp \left(-\frac{1}{2\sigma^2} \mathbf{C}_{\delta}' P_{\delta^c} \mathbf{C}_{\delta} \right),$$

which leads to the strongly consistent Bayes factor estimator

$$(7) \quad \widehat{\text{BF}}_{\delta} = \frac{1}{N} \sum_{i=1}^N (\phi^{2(i)})^{d^c/2} |\mathbf{Z}_{\delta^c}^{(i)'} \mathbf{Z}_{\delta^c}^{(i)}|^{1/2} \times \exp \left(-\frac{1}{2\sigma^{2(i)}} \left(\frac{|\boldsymbol{\gamma}_{\delta^c}^{(i)}|^2}{\phi^{2(i)}} + \mathbf{C}_{\delta}^{(i)'} P_{\delta^c}^{(i)} \mathbf{C}_{\delta}^{(i)} \right) \right).$$

Details and proofs of the results given here are in Supplemental Information, Section D [Li et al. (2011)].

3.3. Increasing computational speed. For data sets with large numbers of SNPs and phenotypes, the slow computation speed of the Gibbs sampler can be a major problem. We have identified two bottlenecks. First, if the number of SNPs is increased, then for each iteration, the number of missing SNPs to be updated will also increase. Second, in the iterations of the Gibbs sampler, the generation of γ involves inverting the matrix $\mathbf{Z}'R^{-1}\mathbf{Z} + (1/\phi^2)I$ each time, as the \mathbf{Z} matrix changes at each iteration. We address these in the following sections.

3.3.1. SNP updating. To speed up calculation, we show that instead of updating all the SNPs at each iteration, updating only one column of SNPs (that is, one SNP updated for all observations) at each cycle will still conserve the target stationary distribution and ergodicity. As the SNP has only three possible values, this change should not have a great effect on the mixing.

A consequence of this result is that instead of updating tens or hundreds of SNPs in one cycle, we need to update just one SNP in each cycle. This single-SNP updating will dramatically speed up computation, especially when there are large numbers of SNPs, or large numbers of observations, in the data. (See in Supplemental Information [Li et al. (2011)], Section E.)

3.3.2. Matrix inverse updating. In the iterations of the Gibbs sampler, a major bottleneck is the generation of γ , since it involves inverting the matrix $\mathbf{Z}'R^{-1}\mathbf{Z} + (1/\phi^2)I$ each time, as the \mathbf{Z} matrix changes at each iteration. Two modifications will speed up this calculation, each based on Woodbury's formula [see Hager (1989) and Appendix C].

By Woodbury's formula, if the matrices A and $I - VA^{-1}U$ are both invertible, then

$$(8) \quad (A + UV)^{-1} = A^{-1} - A^{-1}U(I + VA^{-1}U)^{-1}VA^{-1}.$$

If U and V are vectors, the inverse takes a particularly nice form:

$$(9) \quad (A + uv')^{-1} = A^{-1} - \frac{A^{-1}uv'A^{-1}}{(1 + v'A^{-1}u)},$$

so if we have A^{-1} , no further inversion is needed.

First, relating to the generation of γ in (2), (8) leads to the identity

$$(10) \quad \left(\mathbf{Z}'R^{-1}\mathbf{Z} + \frac{1}{\phi^2}I\right)^{-1} = \phi^2 \left[I - \mathbf{Z}' \left(\frac{1}{\phi^2}R + \mathbf{Z}\mathbf{Z}' \right)^{-1} \mathbf{Z} \right],$$

where the left-hand side involves the inversion of an $s \times s$ matrix, and the right-hand side involves the inversion of an $n \times n$ matrix. Thus, we can always choose to invert the smaller matrix.

Next we look at inverting $\mathbf{Z}'R^{-1}\mathbf{Z} + (1/\phi^2)I$ [a similar argument can be developed for the right-hand side of (10)]. Suppose, at the current iteration, we have $A_0 = \mathbf{Z}'_0R^{-1}\mathbf{Z}_0 + (1/\phi_0^2)I$, and we update to $A_1 = \mathbf{Z}'_1R^{-1}\mathbf{Z}_1 + (1/\phi_1^2)I$. Because we update one column of SNPs at each iteration, we have $\mathbf{Z}_1 = \mathbf{Z}_0 + \Delta$, where Δ is a matrix of all 0's, except for one column. This column contains the differences of the respective columns from \mathbf{Z}_1 and \mathbf{Z}_0 . Thus, $\Delta = (0 \cdots 00\delta 0 \cdots 0)$, and

$$A_1 = A_0 + \Delta'R^{-1}\mathbf{Z}_0 + \mathbf{Z}'_0R^{-1}\Delta + \Delta'R^{-1}\Delta + \left(\frac{1}{\phi_1^2} - \frac{1}{\phi_0^2}\right)I.$$

The three matrices on the right-hand side involving Δ are all rank one matrices, that is, they are of the form uv' for column vectors u and v . Moreover, we can write $I = \sum_{j=1}^s e_j e_j'$, where e_j is a column vector of zeros with a 1 in the j th position. We can then apply (9) three times to get the inverse of A_1 . This calculation involves only matrix by vector multiplications for the middle three terms on the right-hand side. For the e_j vectors, the multiplications reduce to an element extraction. (See Appendix C for details.)

4. Empirical analyses of BAMD.

4.1. *Percentage of missingness handled.* In this subsection we apply BAMD to simulated data in order to assess the procedure's performance as we increase the percentage of missing data in the \mathbf{Z} matrix. We simulated a data set with six families, 20 observations in each family and 5 SNPs per observation. The five SNPs are independent of each other. The six families are also independent, so that the parents of the six families are not related and individuals across families are independent. On the other hand, the individuals within each family share the same parents; this relationship is captured via the numerator relationship matrix. From this data set, four data sets with different percentages of missing values, 5%, 10%, 15%, and 20%, were randomly derived. The family effects, β , which were used to simulate the data, are listed in Table 1. The true SNP effects (additive and dominant effects) used to generate the data are listed in Table 2. When simulating, we let the variance parameter $\sigma^2 = 1$. Our proposed methodology was applied to analyze the data without missing values and also to the new data containing missing values.

Note that for this small simulation, we used a parameterization different from the $\{-1, 0, 1\}$ coding that we use for larger numbers of SNPs. In this example, each SNP effect is represented as (γ_a, γ_d) —the additive and dominant effects of the SNP genotypes.

Tables 1 and 2 summarize the parameter estimation capabilities of BAMD for family and SNP effects. All calculations were based on samples obtained after an initial burn-in of 20,000 iterations of BAMD. The results show that when the percentage of missing values is less than 15%, the proposed

TABLE 1

The true family effects for the simulated data set are given in the first row of the table. The remaining rows indicate the estimated means returned from running BAMD on the data sets derived by setting different degrees of missing values in the SNP matrix of the simulated data set

	β_1	β_2	β_3	β_4	β_5	β_6
Actual value	15	20	25	30	35	40
0% missing	15.45	20.65	25.48	29.84	34.76	40.40
5% missing	15.16	20.74	25.46	28.29	33.43	38.62
10% missing	16.18	21.38	25.65	30.71	35.86	40.81
15% missing	15.45	19.63	24.59	30.18	35.38	40.18
20% missing	14.87	20.18	24.68	30.08	34.88	40.13

methodology yields good estimates for the parameters of direct interest. When the percentage of missing values is greater than 15%, we should be wary of interpreting the results. For example, the true dominant effect for SNP 3 is 0, but the estimate is 1.32 when the percentage of missing values is 20%. Note that the estimate in this case is accurate when the percentage of missing values is less than 10%. We believe the discrepancy arises because one category of genotype for SNP 3 has substantially higher probability and it overpowers the other two categories. When the percentage of missing values increases, the dominated genotype category has only a small chance to be well represented and thus may have unreliable estimates.

Our ultimate goal is to identify significant SNPs from the candidate SNPs. Since we believe that imputation is a tool to obtain better estimates of the parameters, we are not particularly interested in recovering the actual imputed values for the missing SNPs. Nonetheless, the simulation results in Table 3 show that when the probability of one genotype for a certain SNP is dominantly high, the imputed SNPs are correctly identified with probability

TABLE 2

The true additive and dominant effects for each SNP in the simulated data set are given in the first row of the table. The remaining rows indicate the estimated SNP effects returned from running BAMD on the data sets derived by setting different degrees of missing values in the SNP matrix of the simulated data set

	SNP1:a	SNP1:d	SNP2:a	SNP2:d	SNP3:a	SNP3:d	SNP4:a	SNP4:d	SNP5:a	SNP5:d
Actual SNP	-2.00	1.00	1.00	-1.00	3.00	0.00	2.50	0.10	0.30	3.00
0% missing	-2.16	1.00	0.82	-0.75	2.59	0.30	2.43	0.60	-0.04	2.59
5% missing	-1.86	1.14	1.16	-1.05	3.00	0.05	2.21	-0.20	0.48	3.00
10% missing	-1.95	0.77	1.18	-1.52	2.74	0.18	2.51	0.13	0.00	2.74
15% missing	-1.80	0.78	0.99	-0.96	2.48	0.67	2.43	0.47	0.73	2.48
20% missing	-2.08	1.29	1.21	-0.76	3.10	1.32	1.87	-0.20	0.47	3.10

TABLE 3

The true genotype probabilities for the SNPs used to generate the simulated data set are given in the first 3 rows. The final row identifies the frequency with which the true genotype was imputed when running BAMD with 10% of missing data in the SNP matrix

	SNP1	SNP2	SNP3	SNP4	SNP5
	$a = -2$	$a = 1$	$a = 3$	$a = 2.5$	$a = 0.3$
	$d = 1$	$d = -1$	$d = 0.5$	$d = 0.1$	$d = 3$
Actual SNP					
Pr(GG)	0.1309	0.3012	0.8181	0.7719	0.3983
Pr(GC)	0.5307	0.3875	0.0796	0.1950	0.5425
Pr(CC)	0.3384	0.3113	0.1023	0.0331	0.0592
Frequency of correct imputation	0.5500	0.5479	0.6337	0.8508	0.6516

ranging from 0.55–0.85, being correctly imputed more frequently than the other genotypes distributions (see SNPs 3 and 4).

4.2. *Comparison with BIMBAM.* Here we compare our multiple-imputation missing data algorithm with a program called BIMBAM [Servin and Stephens (2007), <http://stephenslab.uchicago.edu/software.html>], which is a popular program among geneticists for association genetics and variable selection with missing data (using single imputation).

BAMD and BIMBAM both propose a two-stage procedure that involves first finding a set of significant SNPs, and then running these significant SNPs through a variable selection procedure that finds the best subset of the significant SNPs that describes the variation in the phenotype. Hence, in this study, BIMBAM and BAMD are assessed through the SNPs they find in the first stage and through the final model they put forth.

For the evaluation, we simulated data from the model given in equation (1). The dimensions of the model were fixed to be $n = 50$, $p = 3$, and $s = 25$ throughout. The three families comprised 16, 17, and 17 individuals, respectively. In addition, the \mathbf{X} and β matrices were fixed. Entries in the \mathbf{Z} matrix took three possible values, mirroring the real-life situation, when they would represent genotypes. Interest lies in discovering the significant coordinates of the γ vector (which corresponds to SNP effects), in the presence of missing values in the \mathbf{Z} matrix.

In the simulation study, three factors—percentage of missing values in \mathbf{Z} , magnitude of γ effects, and the degree of correlation within a family—were varied across different levels. When a particular factor was being investigated, the others were held constant. Here in the main paper, we only present two specific comparisons. The remainder of the results from the simulation study can be found in Supplemental Information [Li et al. (2011)], Section F. In running the study, we simulated several data sets for each case, and ob-

served very consistent results. Hence, in presenting our results, we focus on a single representative data set in each case.

In both of the studies presented here, the γ vector was generated from a multivariate normal, and any values less than 3 in absolute value were set to 0. After generating the \mathbf{Y} responses, 20% of the entries in the \mathbf{Z} matrix were set to missing before being passed to BAMD and BIMBAM.

The first comparison measures the performance of the procedures when an equicorrelation structure (ρ was set to be 0.8) exists within each of the three families. The second comparison presented here aims to see if BAMD turns up many false positives. The γ vector was generated in the same way as earlier, but only the coordinates with the five largest values were kept. The rest were set to 0. In addition, the individuals were assumed to be uncorrelated, that is, $\mathbf{R} = \mathbf{I}_n$.

The results for each comparison are summarized through two diagrams. The first (the upper panels in Figures 1 and 2) display the SNPs that BAMD and BIMBAM found to be significant in the first stage. The lower panels in Figures 1 and 2 display the output from the variable selection procedure.

Each figure shows that BAMD significantly outperforms BIMBAM. In the first example, BIMBAM found only one of the 14 significant SNPs, while BAMD found six. In the second example, there were five significant SNPs, and BIMBAM only found one again, while BAMD found four of them.

5. Analysis of the loblolly pine data. Carbon isotope discrimination (CID) is a time-integrated trait measure of water use efficiency. González-Martínez et al. (2008) used the family-based approach of the Quantitative Transmission Disequilibrium Test QTDT to detect SNPs associated with CID. We utilized the family structure of this population [also described in Kayihan et al. (2005)] in the design matrix in our model. Of the 61 control-pollinated families measured for CID, each had approximately 15 offspring that were clonally propagated by rooted cuttings to generate the ramets (genetically identical replicates). Each genotype has two ramets, sampled from each of two testing sites at Cuthbert, GA and Palatka, FL. Our approach enables us to utilize the family pedigree and parental information to recover missing SNP genotypes. With informative priors, we infer the progeny SNP genotype through Mendelian randomization [Falconer and Macay (1996)]. With uninformative priors, we assume SNPs are missing at random and assign equal probability for each genotype class for missing SNPs.

All SNPs are simultaneously tested under our association model. The Gibbs sampler ran for 50,000 iterations. The first 10,000 iterations were burn-in, after which we thinned the chain every 4 iterations; the autocorrelation reduced significantly after thinning (data not shown). Thus, we have a total of 10,000 samples for each chain for our statistics, and we then applied the variable selector on the four SNPs that were found when using the informative prior.

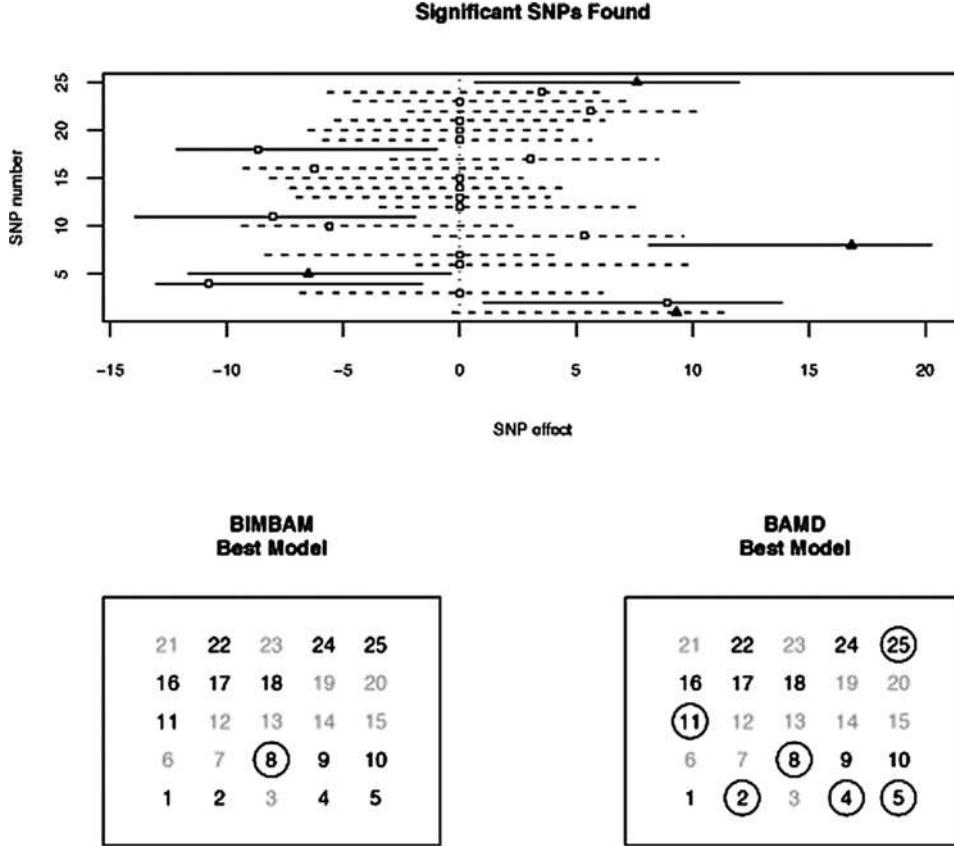


FIG. 1. In the upper panel, triangles and squares represent the true coordinates of the γ vector, where the true nonzero SNPs in the model were (1), (2), (4), (5), (8), (9), (10), (11), (16), (17), (18), (22), (24), and (25). A solid triangle means that BIMBAM found that SNP to be significant at $\alpha = 0.05$ level, the remaining SNPs are squares. Horizontal lines represent highest posterior density intervals returned by BAMD. Solid lines mean the 95% HPD interval found that SNP to be significant. Thus, in the SNP-discovery stage, BIMBAM found SNPs (1), (5), (8), and (25) to be significantly nonzero while BAMD picked out SNPs (2), (4), (5), (8), (11), (18), and (25). In the lower panel, the gray numbers are SNPs that were exactly 0 in the true model, and the black numbers are SNPs with nonzero effects. The circled numbers are the SNPs that were in the best model found by the procedure. Thus, the best model found by BIMBAM contains only SNP (8), whereas the best model found by BAMD contains SNPs (2), (4), (5), (8), (11), and (25).

We detected significant effects of several SNPs on CID at a 95% Bayesian confidence interval (Table 4). Using the uninformative prior, we found 3 significant SNPs [(3) ccoamt_s10, (5) ein2_s1, (31) Caf1_s1]. Using the informative prior, we detected 4 SNPs [(5) ein2_s1, (6) cpk3_s5, (29) dhn1_s2, (31) Caf1_s1] as significant. Note that (6) and (29) are close to significant using the uninformative prior, and (3) was close to significant using the in-

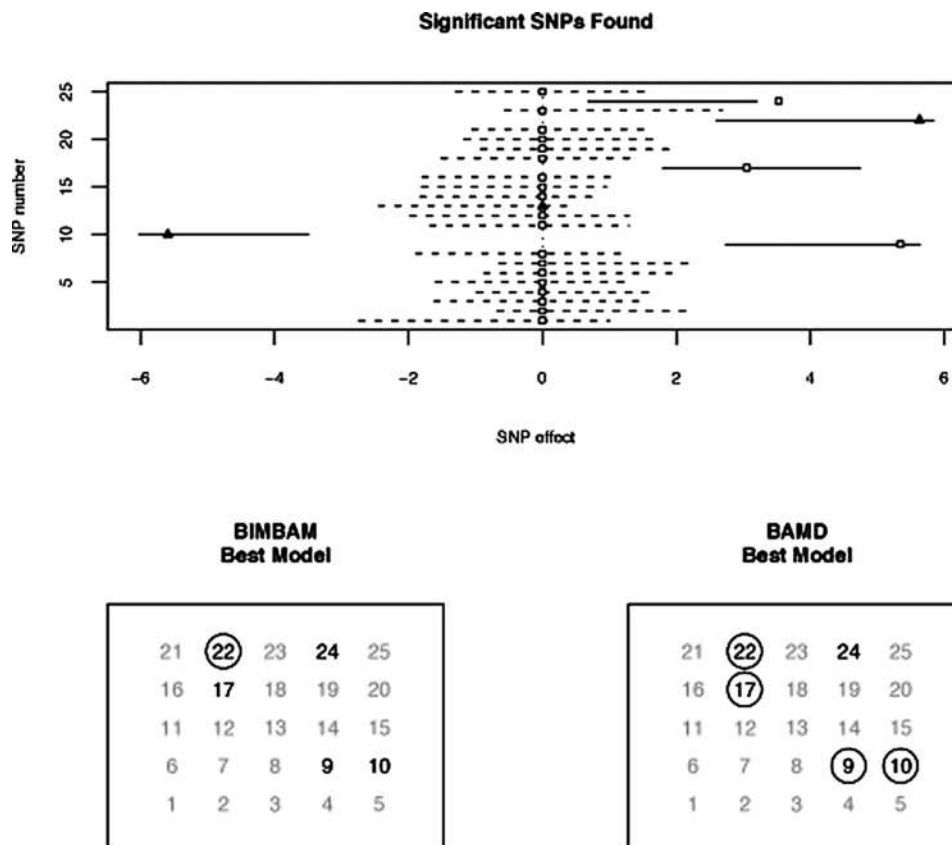


FIG. 2. In the upper panel triangles and squares represent the true coordinates of the γ vector, where the true nonzero SNPs in the model were (9), (10), (17), (22), and (24). A solid triangle means that BIMBAM found that SNP to be significant at $\alpha = 0.05$ level, the remaining SNPs are squares. Horizontal lines represent highest posterior density intervals returned by BAMD. Solid lines mean the 95% HPD interval found that SNP to be significant. Thus, in the SNP-discovery stage, BIMBAM found SNPs (10), (13), and (22) to be significantly nonzero while BAMD found SNPs (9), (10), (17), (22), and (24). In the lower panel, the gray numbers are SNPs that were exactly 0 in the true model, and the black numbers are SNPs with nonzero effects. The circled numbers are the SNPs that were in the best model found by the procedure. The best model found by BIMBAM contains only SNP (22), whereas the best model found by BAMD contains SNPs (9), (10), (17), and (22).

formative prior. This suggests that for these data, the effect of the prior information is important. The QTDT test resulted in 4 significant SNPs, (3), (5), (29), (31), all of which were detected by BAMD which, in addition, found SNP (6). Moreover, it is important to note that BAMD detected these SNPs simultaneously, an indication that their collective effect on the phenotype is being detected.

TABLE 4
Significant SNPs from QTDT tests and the results from the BAMD association model, with 95% confidence intervals

	SNP		Type [‡]
	Informative prior 95% C.I.	Uninformative prior 95% C.I.	
(3) <i>caf1_s1*</i>	(−0.008, 0.110)	(0.013, 0.129)	Syn
(5) <i>ccoamt_s10*†</i>	(−0.103, −0.012)	(−0.097, −0.005)	NC(intron)
(6) <i>cpk3_s5</i>	(−0.052, −0.004)	(−0.048, 0.001)	Syn
(29) <i>dhn1_s2*†</i>	(0.065, 0.113)	(0.044, 0.092)	NC(3′UTR)
(31) <i>ein2_s1*†</i>	(0.077, 0.142)	(0.067, 0.126)	NC(3′UTR)

* Indicates significant in González-Martínez et al. (2008). Bold type indicates significant at the 5% level from our association testing, the rest being nonsignificant. † indicates presence in best model found by variable selector. As indicated in González-Martínez et al. (2008), there are additional SNPs that are marginally significant at $\alpha = 0.1$, which we also detected. ‡: Syn, synonymous SNP; NC, noncoding; UTR, untranslated region.

The use of tag SNPs in a pedigree does not allow for “fine mapping” to SNP effects [Neale and Ingvarsson (2008), Flint-Garcia, Thornsberry and Buckler (2003)]. Thus, the effects of these SNPs on carbon isotope discrimination may reflect the involvement of many linked genes on the phenotype.

We also provide Figures 3 and 4, showing the results for all of the SNPs in the data set. Figure 3 is based on using uninformative SNP priors, while Figure 4 uses informative priors. Although there are few differences in the graphs (showing the strength of the data with respect to the model), we see that the prior can matter. For example, SNP (3) is significant when the noninformative prior is used, but not so when we use the informative prior. The opposite finding holds for SNP (6). Looking at the figures, we see that the significant intervals only barely cross zero; thus, the inclusion of relevant prior information can be quite important.

The four SNPs picked out when using the informative prior were (5), (6), (29), and (31). Due to the small number of variables under consideration, the variable selector procedure was able to run through all 16 possible models. The one with the highest Bayes factor was found to contain SNPs (5), (29), and (31).

6. Discussion. Association testing is being applied to discover relationships among SNPs and complex traits in plants and animals [Flint-Garcia, Thornsberry and Buckler (2003), Hirschhorn and Daly (2005), Balding (2006), Zhu et al. (2008)]. Our model was developed specifically for detecting associations in loblolly pine data, but can be applied to other species as well. Here we discuss some features and limitations of the method.

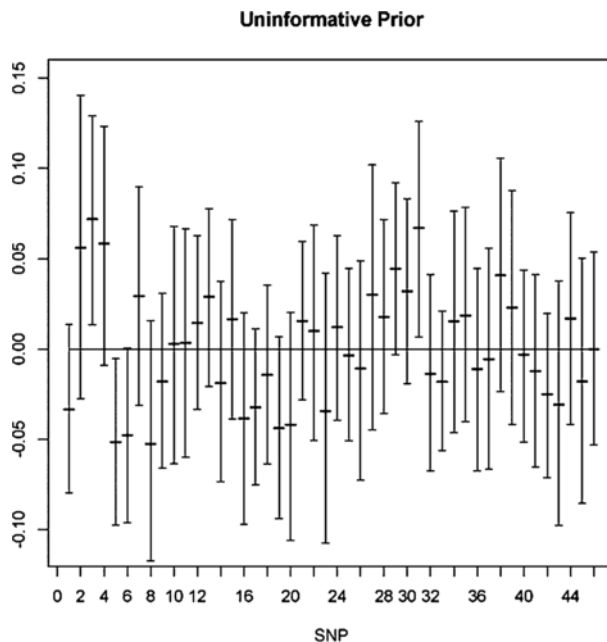


FIG. 3. 95% Confidence intervals for the 44 SNPs from the carbon isotope data, based on 10,000 Gibbs samples from the BAMD model using uninformative priors (equal probability) for the missing SNPs. The significant SNPs are those with intervals that do not cross 0, SNPs (3) *caf1*, (5) *ccoamt*, and (31) *ein2*.

Multiple imputation. Multiple imputation of missing SNP data is the best way to ensure unbiased parameter estimates, which is an important consideration given that SNP effects tend to be small for complex traits of greatest biological interest, and given that results of association studies typically motivate more detailed and labor-intensive investigations of how and why associations were detected.

We used simulation to compare BAMD and BIMBAM for their detection of “correct” vs. “incorrect” SNPs, and found that BAMD performed better than BIMBAM. In practice, this advantage of BAMD over BIMBAM would likely be greatest when missing SNPs are not in LD with nearby SNPs (or adjacency cannot be determined). This is the case in many species, including loblolly pine, in which LD is low and genomic resources such as high-resolution genomic maps and high-density SNP chips for genome scanning are not as well developed as they are for the human genome. The higher computational intensity required for formal multiple imputation in BAMD is a trade-off, however, this has not restricted its practical use in most data sets. For very large data sets, parallel processing seems a logical next step in further increasing the computational efficiency of BAMD.

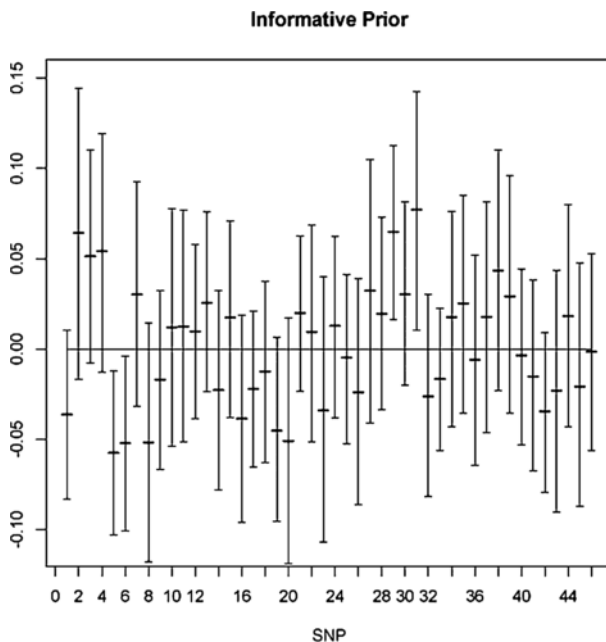


FIG. 4. 95% Confidence intervals for the 44 SNPs from the carbon isotope data, based on 10,000 Gibbs samples from the BAMD model using informative priors (Mendelian randomization) for the missing SNPs. The significant SNPs are those with intervals that do not cross 0, SNPs (5) *ccoamt*, (6) *cpk3*, (29) *dhn1*, and (31) *ein2*.

Family structure. Our method can be applied to family-based association populations, populations of unrelated genotypes, or combination populations. It can incorporate prior information if known. [The application of BAMD in Quesada et al. (2010) was to a population of unrelated genotypes, where significant SNPs related to disease resistance were found.]

Probit models. Although here we assume a continuous response variable, the method can be adapted to discrete phenotypes using a probit link. For example, in a case control study, the response would be either case or control status, and with a probit model we add a latent variable in the Gibbs sampler.

SNP detection. Although BAMD successfully detected the same significant SNPs as were previously detected using the family-based method QTDT [González-Martínez et al. (2008)], as well as an additional significant SNP, the BAMD variable selector indicated that a subset of the significant SNPs was sufficient to explain variation in the phenotype carbon isotope discrimination. This is a useful tool for biologists because a simultaneous solution for SNP effects enables detection of numerous SNPs that collectively

explain phenotypes, which in turn enables further biological experiments to investigate their underlying basis.

However, the candidate SNPs found by BAMD and QTDT cannot necessarily be deemed “correct” or “incorrect” without additional biological experiments. As such, little more can be stated about the correctness of SNPs 3, 5, 29, and 31 without validation experiments. In the broader context of association testing, it is relevant to note that the use of QTDT is limited to families, whereas BAMD and BIMBAM can be used to detect associations in families as well as populations of unrelated individuals. The ability to use BAMD and BIMBAM in many different types of populations is appealing.

Simultaneous vs. genome-wide. Genome-wide association studies are, typically, marginal processors of the data. That is, each SNP is assessed individually for its association, so simultaneous assessment, or epistasis, cannot be detected. A model such as (1) is assessing the SNPs simultaneously—that is its strength. But how many SNPs should we expect to be able to handle in one model? Computational issues aside, if the number of SNPs is greatly increased, we are then susceptible to the usual regression pitfalls—multicollinearity being the most prevalent. Thus, we recommend using BAMD on smaller sets of SNPs that have had some preprocessing. Thus far, BAMD has been used successfully on a model with 400 SNPs [Quesada et al. (2010)], and we have tested it on as many as 800 SNPs.

Missing data. The missing data problem is common across all genomics data sets, so there is broad potential utility of this method. The assumption of MAR (missing at random), which is reasonable in these contexts, may bear additional research attention. If there are quality concerns about SNP data, there are some statistical steps forward, as noted by Wilson et al. (2010), such as using indicator variables of missingness as predictors. This approach can even be extended to test if missingness is a heritable trait and, if so, the MAR assumption is invalid. Next generation sequencing platforms may generate sufficient data to enable this assumption to be tested and, if borne out, may motivate placement of priors on SNP calls in certain sequence contexts.

Last, software to run the Gibbs sampler and variable selector is available in the R package BAMD.

APPENDIX A: CALCULATING THE NUMERATOR RELATIONSHIP MATRIX

The algorithm is due to Henderson (1976) and Quaas (1976). The individuals within 61 families and the parents for the 61 families are ordered together such that the first $1, \dots, a$ subjects are unrelated and are used as

a “base” population. Let the total number of subjects within families and parents of the 61 families be n , and we will get a numerator relationship matrix with dimension $n \times n$. As the first a subjects (being part of the parents of the 61 families) are unrelated, the upper left submatrix with dimension $a \times a$ of the numerator relationship matrix is identity matrix I . This identity submatrix will be expanded iteratively until it reaches to dimension $n \times n$.

As we know the sub-numerator relationship matrix for the first unrelated a subjects is the identity, next we will give the details how to calculate the remaining cells of the numerator relationship matrix for the related subjects. Consider the j th and the i th subject from the above ordered subjects:

- (1) If both parents of the j th individual are known, say, g and h , then

$$R_{ji} = R_{ij} = 0.5(R_{ig} + R_{ih}), \quad i = 1, \dots, j-1;$$

$$R_{jj} = 1 + 0.5R_{gh},$$

where R_{ji} is the cell of the numerator relationship matrix in the j th row and i th column.

- (2) If only one parent is known for the j th subject, say, it is g , then

$$R_{ji} = R_{ij} = 0.5R_{ig}, \quad i = 1, \dots, j-1;$$

$$R_{jj} = 1.$$

- (3) If neither parent is known for the j th subject,

$$R_{ji} = R_{ij} = 0, \quad i = 1, \dots, j-1;$$

$$R_{jj} = 1.$$

For the loblolly pine data, we have 44 pines acting as grandparents and they produce 61 pine families. The 61 families contains 888 individual pine trees all together, also called clones. The phenotypic responses are taken from the individual clones. So our interest is in calculating the relationship matrix for the 888 clones and it would have a dimension 888×888 . According to Henderson’s method, we ordered the 44 grandparent pines and 888 individual pines together such that the first a pines are not related. Starting from the $(a+1)$ th pine, we applied the above iteration calculation algorithm, and in the end had a relationship matrix with dimension 932×932 for all the grandparent pines and all individual clones. We took a submatrix from the right bottom of the previous numerator relationship matrix with dimension 888×888 and it is the numerator relationship matrix we used in the loblolly pine data analysis.

APPENDIX B: ESTIMATION WITH THE EM ALGORITHM

B.1. Missing data. The EM algorithm begins by building the *complete data likelihood*, which is the likelihood function that would be used *if the missing data were observed*. When we fill in the missing data we write $Z_i^* =$

(Z_i^o, Z_i^m) , and the *complete data* are (Y, Z^*) with likelihood function

$$(11) \quad L_C \propto \prod_{i \in I_0} \exp\left(-\frac{1}{2\sigma^2}(Y_i - X_i\beta - Z_i\gamma)^2\right) \\ \times \prod_{i \in I_M} \exp\left(-\frac{1}{2\sigma^2}(Y_i - X_i\beta - Z_i^*\gamma)^2\right),$$

where I_0 indexes those individuals with complete SNP data, and I_M indexes those individuals with missing SNP information.

The *observed data likelihood*, which is the function that we eventually use to estimate the parameters, must be summed over all possible values of the missing data. So we have

$$L_o \propto \frac{1}{(2\pi\sigma^2)^{n/2}} \prod_{i \in I_0} \exp\left(-\frac{1}{2\sigma^2}(Y_i - X_i\beta - Z_i\gamma)^2\right) \\ \times \prod_{i \in I_M} \sum_{Z_i^*} \exp\left(-\frac{1}{2\sigma^2}(Y_i - X_i\beta - Z_i^*\gamma)^2\right).$$

The distribution of the missing data Z_i^* is given by the ratio of L_C/L_o :

$$(12) \quad P(Z_i^*) = \frac{\exp(-(Y_i - X_i\beta - Z_i^*\gamma)^2/(2\sigma^2))}{\sum_{Z_i^*} \exp(-(Y_i - X_i\beta - Z_i^*\gamma)^2/(2\sigma^2))},$$

where the sum in the denominator is over all possible realizations of Z_i^* . This is a discrete distribution on the missing SNP data for each individual. To understand it, look at one individual.

Suppose that there are g possible genotypes (typically $g = 2$ or 3) and individual i has missing data on k SNPs. So the data for individual i is $Z_i = (Z^o, Z^m)$, where Z^m has k elements, each of which could be one of g classes. For example, if $g = 3$ and $k = 7$, then Z^m can take values in the following:

		SNP						
Genotype	*				*			
		*	*			*		
				*				*

where the * show one possible value of the Z_i^m . For the example, there are $3^7 = 2187$ possible values for Z_i^m . In a real data set this could grow out of hand. For example, if there were 12 missing SNPs, then there are 531,441 possible values for Z_i^m ; with 20 missing SNPs the number grows to 3,486,784,401 (3.5 billion).

B.2. An EM algorithm. To the expected value of the log of the complete data likelihood (11), we only deal with the second term (with the missing

data). This expected value does not change the piece with no missing data, but does change the second piece. Standard calculations give

$$\mathbb{E}\left(\frac{1}{2\sigma^2}(Y_i - X_i\beta - Z_i^*\gamma)^2\right) = \frac{1}{2\sigma^2}(Y_i - X_i\beta - \mathbb{E}(Z_i^*\gamma))^2 + \text{Var}(Z_i^*\gamma),$$

where

$$(13) \quad \mathbb{E}(Z_i^*) = (Z_i^o, \mathbb{E}(Z_i^m)) \quad \text{and} \quad \text{Var}(Z_i^*\gamma) = \text{Var}(Z_i^m\gamma_i^m).$$

If we define

$$Y_{n \times 1} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad X_{n \times p} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}, \quad \mathbf{Z}_{n \times s} = \begin{pmatrix} (Z_1^o, \mathbb{E}(Z_1^m)) \\ (Z_2^o, \mathbb{E}(Z_2^m)) \\ \vdots \\ (Z_n^o, \mathbb{E}(Z_n^m)) \end{pmatrix},$$

the expected complete data log likelihood is

$$(14) \quad \mathbb{E} \log L_C = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} |Y - X\beta - \mathbf{Z}\gamma|^2 - \frac{1}{2\sigma^2} \gamma' V_Z \gamma,$$

where V_{Z_i} is the variance-covariance matrix of the vector Z_i with elements given by

$$V_{Z_{i_j j'}} = \begin{cases} 0, & \text{if either } Z_{i_j} \text{ or } Z_{i_{j'}} \text{ is observed,} \\ \text{Cov}(Z_{i_j}, Z_{i_{j'}}), & \text{if neither } Z_{i_j} \text{ nor } Z_{i_{j'}} \text{ is observed,} \end{cases}$$

and $V_Z = \sum_{i \in I_M} V_{Z_i}$. Standard calculus will show that the MLEs from (14) are given by

$$(15) \quad \begin{aligned} \hat{\beta} &= (X'X)^{-1} X'(Y - \mathbf{Z}\hat{\gamma}), \\ \hat{\gamma} &= (\mathbf{Z}'\mathbf{Z} - V_Z)^{-1} \mathbf{Z}'(I - H)Y, \\ \hat{\sigma}^2 &= \frac{1}{n} (|Y - X\hat{\beta} - \mathbf{Z}\hat{\gamma}|^2 + \hat{\gamma}' V_Z \hat{\gamma}). \end{aligned}$$

The algorithm now iterates between (12), (13), and (15) until convergence.

B.3. Implementation. To implement the EM algorithm, we must be able to either:

- (1) calculate the expectation and variance in (13), or
- (2) generate a random sample from (12) and calculate the terms in (13) by simulation.

The first option is impossible and the second is computationally intensive, but the only way.

Going back to (12), note that this is the distribution of the vector of missing values for individual i . If the data are $Z_i^* = (Z_i^o, Z_i^m)$, we are only

concerned with $Z_i^m = (Z_{i1}^m, \dots, Z_{ik}^m)$, and for $\mathbf{c} = (c_1, \dots, c_k)$,

$$P(Z_i^m = \mathbf{c}_0) = \frac{\exp(-(Y_i - X_i\beta - Z_i^o\gamma_i^o - \mathbf{c}_0\gamma_i^m)^2/(2\sigma^2))}{\sum_{\text{all } \mathbf{c}} \exp(-(Y_i - X_i\beta - Z_i^o\gamma_i^o - \mathbf{c}\gamma_i^m)^2/(2\sigma^2))},$$

where the sum in the denominator can easily have over 1 billion terms.

A possible alternative is to use a Gibbs sampler to simulate the distribution of Z_i^m by calculating the distribution of each element conditional on the rest of the vector. For a particular element Z_{ij}^m , the conditional distribution given the rest of the vector $Z_{i(-j)}^m$ is given in (3). So to produce a sample of Z_i^m , we loop through a Gibbs sampler.

Unfortunately, there may be problems with this algorithm in that it may still be too computationally intensive. The Gibbs samplers (2) and (3) need to be run for every iteration of the EM algorithm. For each iteration of EM we may need 20–50 thousand Gibbs iterations. If there is a lot of missing data, this could result in a very slow algorithm.

APPENDIX C: MATRIX INVERSE UPDATES

We are interested in matrices of the form $A_0 + \sum_{k=1}^p u_k v_k'$, where $u_k, v_k, k = 1, \dots, p$, are vectors. For this form we have the following lemma, which follows from Woodbury's formula.

LEMMA 1. *Let A_0 be invertible, and $u_j, v_j, j = 1, \dots, p$, be vectors. Define*

$$A_k = A_0 + \sum_{j=1}^k u_j v_j'.$$

Then for $k = 1, \dots, p$,

$$(16) \quad A_k^{-1} = A_{k-1}^{-1} - \frac{A_{k-1}^{-1} u_k v_k' A_{k-1}^{-1}}{1 + v_k' A_{k-1}^{-1} u_k}.$$

Then, to calculate $A_p^{-1} = (A_0 + \sum_{k=1}^p u_k v_k')^{-1}$, we can start with A_0^{-1} , and use the recursion to get to A_p^{-1} . Note that each step of the recursion requires only multiplication of matrices by vectors. Moreover, in many applications the vectors u_k, v_k are sparse, so the multiplication amounts to extracting elements.

SUPPLEMENTARY MATERIAL

Theory and additional simulations (DOI: [10.1214/11-AOAS516SUPP](https://doi.org/10.1214/11-AOAS516SUPP); .pdf). The Supplemental Information contains details on the variable selector, and the proof of convergence of the two Markov chains (the Gibbs sampler and the model search). In addition, there are further comparisons between BAMD and BIMBAM.

REFERENCES

- BALDING, D. J. (2006). A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.* **7** 781–791.
- CASELLA, G., GIRÓN, F. J., MARTÍNEZ, M. L. and MORENO, E. (2009). Consistency of Bayesian procedures for variable selection. *Ann. Statist.* **37** 1207–1228. [MR2509072](#)
- CHEN, W. M. and ABECASIS, G. R. (2007). Family-based association tests for genomewide association scan. *The American Journal of Human Genetics* **81** 913–926.
- CHEN, M.-H. and SHAO, Q.-M. (1997). Estimating ratios of normalizing constants for densities with different dimensions. *Statist. Sinica* **7** 607–630. [MR1467451](#)
- DAI, J. Y., RUCZINSKI, I., LEBLANC, M. and KOOPERBERG, C. (2006). Imputation methods to improve inference in SNP association studies. *Genet. Epidemiol.* **30** 690–702.
- FALCONER, D. S. and MACAY, T. F. C. (1996). *Introduction to Quantitative Genetics*, 4th ed. Longman, Harlow.
- FLINT-GARCIA, S. A., THORNSBERRY, J. M. and BUCKLER, E. S. (2003). Structure of linkage disequilibrium in plants. *Annu. Rev. Plant Bio.* **54** 357–374.
- GONZÁLEZ-MARTÍNEZ, S. C., ERSOZ, E., BROWN, G. R., WHEELER, N. C. and NEALE, D. B. (2006). DNA sequence variation and selection of tag single-nucleotide polymorphisms at candidate genes for drought-stress response in *Pinus taeda* L. *Genetics* **172** 1915–1926.
- GONZÁLEZ-MARTÍNEZ, S. C., HUBER, D. A., ERSOZ, E., DAVIS, J. M. and NEALE, D. B. (2008). Association genetics in *Pinus taeda* L. II. Carbon isotope discrimination. *Heredity* **101** 19–26.
- GREENLAND, S. and FINKLE, W. D. (1995). A critical look at methods for handling missing covariates in epidemiologic regression analyses. *Am. J. Epidemiol.* **142** 1255–1264.
- HAGER, W. W. (1989). Updating the inverse of a matrix. *SIAM Rev.* **31** 221–239. [MR0997457](#)
- HENDERSON, C. R. (1976). A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics* **32** 69–83.
- HIRSCHHORN, J. N. and DALY, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Genetics* **6** 95–108.
- HOBERT, J. P. and CASELLA, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *J. Amer. Statist. Assoc.* **91** 1461–1473. [MR1439086](#)
- HUISMAN, M. (2000). Imputation of missing item responses: Some simple techniques. *Quality and Quantity* **34** 331–351.
- KAYIHAN, G. C., HUBER, D. A., MORSE, A. M., WHITE, T. T. and DAVIS, J. M. (2005). Genetic dissection of fusiform rust and pitch canker disease traits in loblolly pine. *Theory of Applied Genetics* **110** 948–958.
- LI, Z., GOPAL, V., LI, X., DAVIS, J. M. and CASELLA, G. (2011). Supplement to “Simultaneous SNP identification in association studies with missing data.” DOI:[10.1214/11-AOAS516SUPP](#).
- LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley, New York.
- MARCHINI, J., HOWIE, B., MYERS, S., McVEAN, G. and DONNELLY, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39** 906–913.
- McKEEVER, D. B. and HOWARD, J. L. (1996). Value of timber and agricultural products in the United States 1991. *Forest Products Journal* **46** 45–50.
- MENG, X.-L. and WONG, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statist. Sinica* **6** 831–860. [MR1422406](#)

- NEALE, D. B. and INGVARSSON, P. K. (2008). Population, quantitative and comparative genomics of adaptation in forest trees. *Curr. Opin. Plant Biol.* **11** 149–155.
- O’HAGAN, A. and FORSTER, J. (2004). *Kendall’s Advanced Theory of Statistics: Vol. 2B: Bayesian Inference*. Arnold, London.
- PRITCHARD, J. K., STEPHENS, M. and DONNELLY, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155** 945–959.
- QUAAS, R. L. (1976). Computing the diagonal elements and inverse of a large numerator relationship matrix. *Biometrics* **46** 949–953.
- QUESADA, T., GOPAL, V., CUMBIE, W. P., ECKERT, A. J., WEGRZYN, J. L., NEALE, D. B., GOLDFARB, B., HUBER, D. A., CASELLA, G. and DAVIS, J. M. (2010). Association mapping of quantitative disease resistance in a natural population of Loblolly pine (*Pinus taeda* L.). *Genetics* **186** 677–686.
- SCHEET, P. and STEPHENS, M. A. (2006). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *Am. Journ. Hum. Genetics* **78** 629–644.
- SERVIN, B. and STEPHENS, M. (2007). Imputation-based analysis of association studies: Candidate regions and quantitative traits. *PLoS Genet.* **3** e114.
- STEPHENS, M., SMITH, N. J. and DONNELLY, P. (2001). A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68** 978–989.
- SU, S. Y., WHITE, J., BALDING, D. J. and COIN, L. J. M. (2008). Inference of haplotypic phase and missing genotypes in polyploid organisms and variable copy number genomic regions. *BMC Bioinformatics* **9** Art. 513.
- SUN, Y. V. and KARDIA, S. L. R. (2008). Imputing missing genotypic data of single-nucleotide polymorphisms using neural networks. *European Journal of Human Genetics* **16** 487–495.
- SZATKIEWICZ, J. P., BEANE, G. L., DING, Y., HUTCHINS, L., DE VILLENA, F. P. and CHURCHILL, G. A. (2008). An imputed genotype resource for the laboratory mouse. *Mammalian Genome* **19** 199–208.
- VAN DER HEIJDEN, G. J., DONDERS, A. R., STIJNEN, T. and MOONS, K. G. (2006). Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example. *J. Clin. Epidemiol.* **59** 1102–1109.
- WEAR, D. N. and GREIS, J. G. (2002). Southern forest resource assessment: Summary of findings. *Journal of Forestry* **100** 6–14.
- WILSON, M. A., IVERSEN, E. S., CLYDE, M. A., SCHMIDLER, S. C. and SCHILDKRAUT, J. M. (2010). Bayesian model search and multilevel inference for SNP association studies. *Ann. Appl. Stat.* **4** 1342–1364. [MR2758331](#)
- YU, J. M., PRESSOIR, G., BRIGGS, W. H., BI, I. V., YAMASAKI, M., DOEBLEY, J., MCMULLEN, M. D., GAUT, B. S., NIELSEN, D. M., HOLLAND, J. B., KRESOVICH, S. and BUCKLER, E. S. (2006). A unified mixed model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* **38** 203–208.
- ZHU, C., GORE, M., BUCKLER, E. S. and YU, J. (2008). Status and prospects of association mapping in plants. *The Plant Genome* **1** 5–20.

Z. LI
STATE STREET CORPORATION
1 LINCOLN STREET, 15TH FLOOR
BOSTON, MASSACHUSETTS 02111
USA

V. GOPAL
DEPARTMENT OF STATISTICS
UNIVERSITY OF FLORIDA
GAINESVILLE, FLORIDA 32611
USA

X. LI
J. M. DAVIS
SCHOOL OF FOREST RESOURCES
AND CONSERVATION
UNIVERSITY OF FLORIDA
GAINESVILLE, FLORIDA 32611
USA

G. CASELLA
DEPARTMENT OF STATISTICS
AND GENETICS INSTITUTE
UNIVERSITY OF FLORIDA
GAINESVILLE, FLORIDA 32611
USA