Simultaneous SNP Identification

in Association Studies with Missing Data

Zhen Li[*]   Vikneswaran Gopal[†]   Xiaobo Li[‡]   John M. Davis[§]

George Casella[¶]

Abstract

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.

Keywords and phrases: Hierarchical models, Bayes models, Gibbs sampling, Genome-wide association

---

[*]State Street Corporation, 1 Lincoln street, 15th floor Boston, MA, 02111

[†]Department of Statistics, University of Florida, Gainesville, FL 32611

[‡]School of Forest Resources and Conservation, University of Florida, Gainesville, FL 32611

[§]School of Forest Resources and Conservation, University of Florida, Gainesville, FL 32611

[¶]Department of Statistics and Genetics Institute, University of Florida, Gainesville, FL 32611

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 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 *et al.* 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 *et al.* 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 1987; Dai *et al.* 2006), particularly in species or genomic regions where LD decays rapidly and thus adjacent SNPs are not necessarily correlated (Flint-Garcia *et al.* 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 *et al.* 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 *et al.* 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 1977), which avoids complications arising from non-positive 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 *et al.* 2001; Servin and Stephens M 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 tag SNP in addition to three 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, and there is an Appendix with technical details and supplemental information.

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 only among SNPs are imputed, 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. To quantify the effect of the SNPs, we consider that each SNP has an additive effect, although the method can be adapted to quantifying additive and dominance effects of SNPs.

We begin with the linear mixed model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \tag{1}$$

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

For our application here (Carbon Isotope Data), we have $n = 1000$ and $s = 450$. In another application of BAMD (Queseda *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 $p >> n$ case.

Each row of the matrix $\boldsymbol{Z}$, $\boldsymbol{Z}_i, i = 1, \ldots n$, 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 $\boldsymbol{Z}_i = (\boldsymbol{Z}_i^{\text{obs}}, \boldsymbol{Z}_i^{\text{miss}})$, where $\boldsymbol{Z}_i^{\text{obs}}$ are the observed genotypes for the *ith* individual, and $\boldsymbol{Z}_i^{\text{miss}}$ are the missing genotypes. Note two things:

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

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

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

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 $\boldsymbol{\gamma} \sim N(\boldsymbol{0}, \sigma^2 \phi^2 \boldsymbol{I}_s)$. 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: $\sigma^2 \sim IG(a, b)$ and $\phi^2 \sim IG(c, d)$, 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 most difficult to justify, is *Missing Completely At Random(MCAR)*. This assumes that 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, $\boldsymbol{R}$, 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

$$\beta \sim N\left(\left(X'R^{-1}X\right)^{-1} X'R^{-1}(Y - Z\gamma), \sigma^2\left(X'R^{-1}X\right)^{-1}\right) \tag{2}$$

$$\gamma \sim N\left(\left(Z'R^{-1}Z + \frac{I}{\phi^2}\right)^{-1} Z'R^{-1}(Y - X\beta), \sigma^2\left(Z'R^{-1}Z + \frac{I}{\phi^2}\right)^{-1}\right)$$

$$\sigma^2 \sim \frac{1}{(\sigma^2)^{n/2+s/2+a+1}} \exp\left(-\frac{(Y - X\beta - Z\gamma)' R^{-1} (Y - X\beta - Z\gamma) + \frac{|\gamma|^2}{\phi^2} + 2b}{2\sigma^2}\right)$$

$$\phi^2 \sim \frac{1}{(\phi^2)^{\frac{s}{2}+c+1}} \exp - \frac{(\frac{|\gamma|^2}{\sigma^2} + 2d)}{2\phi^2}.$$

The SNPs are contained in the 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 jth missing SNP in the ith individual, the conditional distribution given the rest of the vector $Z_{i(-j)}^m$ and all other parameters in the model is

$$P(Z_{ij}^m = c|Z_{i(-j)}^m) = \frac{\exp\left(-\frac{1}{2\sigma^2}(Y_i - X_i\beta - Z_i^o\gamma_i^o - Z_{i(-j)}^m\gamma_{i(-j)}^m - c\gamma_{ij}^m)^2\right)}{\sum_{\ell=1}^3 \exp\left(-\frac{1}{2\sigma^2}(Y_i - X_i\beta - Z_i^o\gamma_i^o - Z_{i(-j)}^m\gamma_{i(-j)}^m - c_\ell\gamma_{ij}^m)^2\right)}. \tag{3}$$

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 E.

3 Variable Selection

The Gibbs sampler will estimate the full set of parameters in model (1). However, it is often the case that there is interest in finding a smaller set of SNPs that will explain a sufficient proportion of the variability, and 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 Z are deleted, giving Z_δ, and the reduced model is

$$Y = X\beta + Z_\delta\gamma_\delta + \varepsilon. \tag{4}$$

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 $\boldsymbol{\theta} := (\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, \phi^2, \boldsymbol{Z})$, and naturally, $\boldsymbol{\theta}_\delta := (\boldsymbol{\beta}, \boldsymbol{\gamma}_\delta, \sigma^2, \phi^2, \boldsymbol{Z}_\delta)$. Let m_δ, π_δ and p_δ denote the marginal distribution of Y, the prior distribution on $\boldsymbol{\theta}_\delta$, and the conditional distribution of 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

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

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 $0 \to 1$ or $1 \to 0$) with probability a and, with probability $1 - a$ we do an independent jump. This is a symmetric candidate, and δ' is accepted with probability $\min\{1, BF_{\delta'}/BF_\delta\}$.

3.2 Estimating the Bayes Factor

Calculating the Bayes factor in (5) requires knowing the $\boldsymbol{Z}$ matrix, which is not the case when there is missing data. Thus, to calculate the Bayes Factor we need to use the imputed Z matrix from the Gibbs sampler. Thus, we run two Markov chains simultaneously:

1. A Gibbs sampler on the full model, to fill in the missing data in the $\boldsymbol{Z}$ and estimate all parameters

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

The aim is to search for δ^* such that $\delta^* = \arg\max_\delta BF_\delta$, 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 in the following proposition.

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

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

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

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

which leads to the strongly consistent Bayes factor estimator

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

Details and proofs of the results given here are in Appendix B.

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. There are two bottlenecks that we have identified and addressed. 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 $Z'R^{-1}Z + (1/\phi^2)I$ each time, as the Z matrix changes at each iteration. We address theses 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) each cycle, will still conserve the target stationary distribution and ergodicity. As the SNP only has 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 just need to update one SNP in each cycle. This will dramatically speed

up computation, especially when there are large numbers of SNPs, or large numbers of observations, in the data. (See Appendix C.)

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 $Z'R^{-1}Z + (1/\phi^2)I$ each time, as the Z matrix changes at each iteration. There are two things that we can do to speed up this calculation, each based on Woodbury's formula (see Hager 1989 and Appendix D).

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

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

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

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

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

First, relating to the generation of γ in (2), From (8) we get the identity

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

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

Next we look at how to invert $Z'R^{-1}Z + (1/\phi^2)I$ (a similar argument can be developed for the right side of (10)). Suppose, at the current iteration, we have $A_0 = Z_0'R^{-1}Z_0 + (1/\phi_0^2)I$, and we update to $A_1 = Z_1'R^{-1}Z_1 + (1/\phi_1^2)I$. Because we update one column of SNPs at each iteration, we have $Z_1 = Z_0 + \Delta$, where Δ is a matrix of all 0s, except for one column. This column contains the differences of the respective columns from Z_1 and Z_0. Thus $\Delta = (0 \cdots 0\ 0\ \delta\ 0 \cdots 0)$, and

$$A_1 = A_0 + \Delta'R^{-1}Z_0 + Z_0'R^{-1}\Delta + \Delta'R^{-1}\Delta + \left(\frac{1}{\phi_1^2} - \frac{1}{\phi_0^2}\right)I.$$

The three matrices on the right 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 jth position. We can then apply (9) three times to get the inverse of A_1. This calculation only involves matrix by vector multiplications for the middle three terms on the right side. For the e_j vectors, the multiplications reduce to an element extraction. (See Appendix D 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 obtain some idea about how the procedure behaves as we increase the percentage of missing data in the Z matrix. We simulated a data set with 6 families, 20 observations in each family and 5 SNPs per observation. The 5 SNPs are independent of each other. The six families are also independent, so that the parents of the 6 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.

A point to note is 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

Table 1: THE TRUE FAMILY EFFECTS FOR THE SIMULATED DATASET 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 DATASET.

Actual Value	β_1	β_2	β_3	β_4	β_5	β_6
	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

Table 2: THE TRUE ADDITIVE AND DOMINANT EFFECTS FOR EACH SNP IN THE SIMULATED DATASET 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 DATASET.

Actual SNP	SNP1:a	SNP1:d	SNP2:a	SNP2:d	SNP3:a	SNP3:d	SNP4:a	SNP4:d	SNP5:a	SNP5:d
	-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

and SNP effects. All calculations were based on samples obtained after an initial burn-in of 20000 iterations of BAMD. The results show that when the percentage of missing values is less than 15%, the proposed methodology yields good estimates for the parameters we are interested in. When the percentage of missing values is greater than that, we should be wary of interpreting the results. Take SNP 3, 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%. The reason for that, we believe, is that one category of genotype for SNP 3 has substantially higher probability and it overpowers the other two. When the percentage of missing values increases, the dominated genotype category has only a small chance to be well represented

Table 3: THE ABOVE TABLE DISPLAYS THE TRUE GENOTYPE PROBABILITIES FOR THE SNPS USED TO GENERATE THE SIMULATED DATASET 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.

Actual SNP	SNP1 $a = -2, d = 1$	SNP2 $a = 1, d = -1$	SNP3 $a = 3, d = 0.5$	SNP4 $a = 2.5, d = 0.1$	SNP5 $a = 0.3, d = 3$
$\Pr(GG)$	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	0.55004	0.54788	**0.63372**	**0.85075**	0.65159

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. That being said, however, 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 a substantially high probability. The true probabilities of the SNPs are presented; those with genotypes that have a high probability are boldfaced. It is apparent that these are also imputed more frequently than the other possible genotypes for that SNP (see SNPs 3 and 4).

4.2 Comparison with BIMBAM

Here we compare our multiple-imputation missing data algorithm with a popular (in practice), single-imputation program called BIMBAM (Servin and Stephens 2007). It is the latest program from the Stephens Lab (http://stephenslab.uchicago.edu/software.html). It is a popular program in the genetics community, doing association genetics and variable selection with missing data, using a single imputation of the missing data.

BAMD and BIMBAM both propose a two-stage procedure that involves first finding a set of significant SNPs, and then running these through a variable selection procedure that finds the best subset of the significant SNPs that describes the variation in the phenotype.

Figure 1: In the upper panel, triangles and squares represent the true co-ordinates of the γ vector, where the true non-zero 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 non-zero 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 non-zero 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).
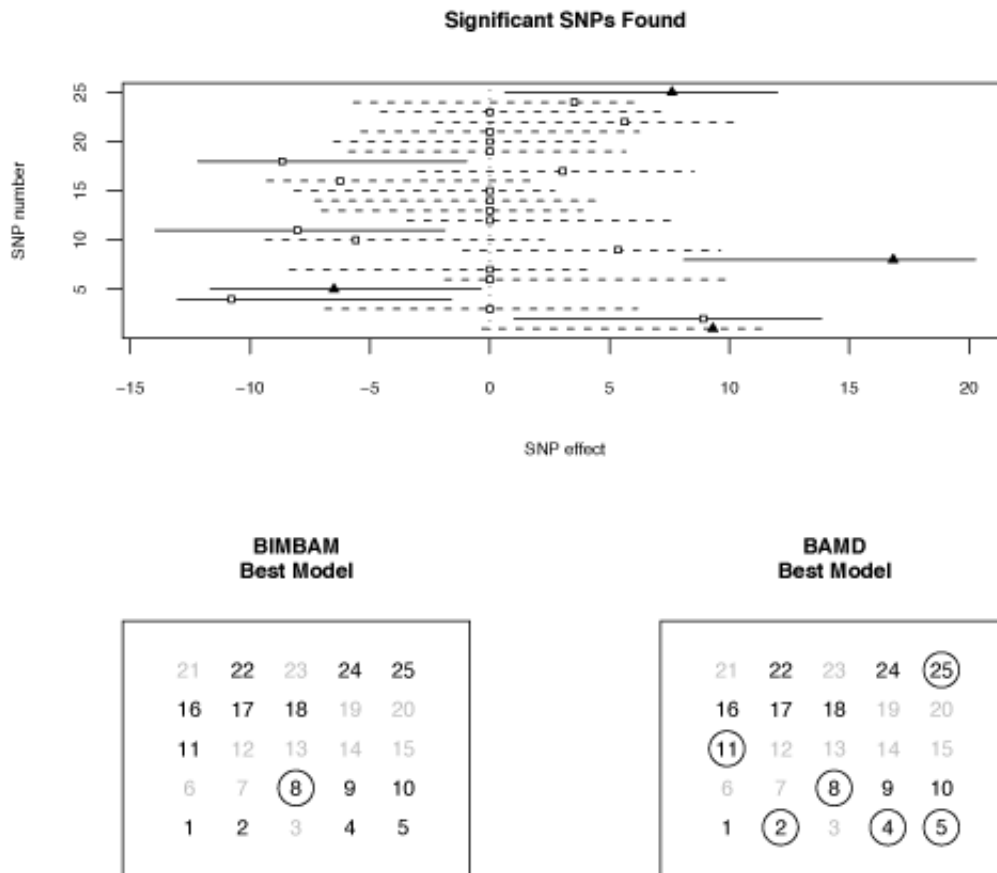
Hence in this study, BIMBAM and BAMD are assessed through the SNPs they find in the first stage and through the final model they put forth.

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 3 families comprised 16, 17 and 17 individuals respectively. In addition, the $\boldsymbol{X}$ and $\boldsymbol{\beta}$ matrices were fixed. Entries in the $\boldsymbol{Z}$ matrix took 3 possible values, mirroring the real-life situation, when they would represent genotypes. Interest lies in discovering the significant co-ordinates of the $\boldsymbol{\gamma}$ vector (which corresponds to SNP effects), in the presence of missing values in the $\boldsymbol{Z}$ matrix.

In the simulation study, 3 factors - percentage of missing values in $\boldsymbol{Z}$, magnitude of $\boldsymbol{\gamma}$ 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 2 specific comparisons. The remainder of the results from the simulation study can be found in Supplemental Information F. In running the study, we simulated several datasets for each case, and observed very consistent results. Hence in presenting our results, we focus on a single representative dataset in each case.

In both of the studies presented here, the $\boldsymbol{\gamma}$ vector was generated from a multivariate normal, and any values less than 3 in absolute value were set to 0. After generating the $\boldsymbol{Y}$ responses, 20% of the entries in the $\boldsymbol{Z}$ 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 3 families. The second comparison presented here aims to see if BAMD turns up many false positives. The $\boldsymbol{\gamma}$ vector was generated in the same way as earlier, but only the co-ordinates with the 5 largest values were kept. The rest were set to 0. In addition, the individuals were assumed to be uncorrelated, that is, $\boldsymbol{R} = \boldsymbol{I}_n$.

The results for each comparison are summarized through 2 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 Figure 1 and 2 display the output from the variable selection procedure.

Figure 2: IN THE UPPER PANEL TRIANGLES AND SQUARES REPRESENT THE TRUE CO-ORDINATES OF THE γ VECTOR, WHERE THE TRUE NON-ZERO 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 NON-ZERO 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 NON-ZERO 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).
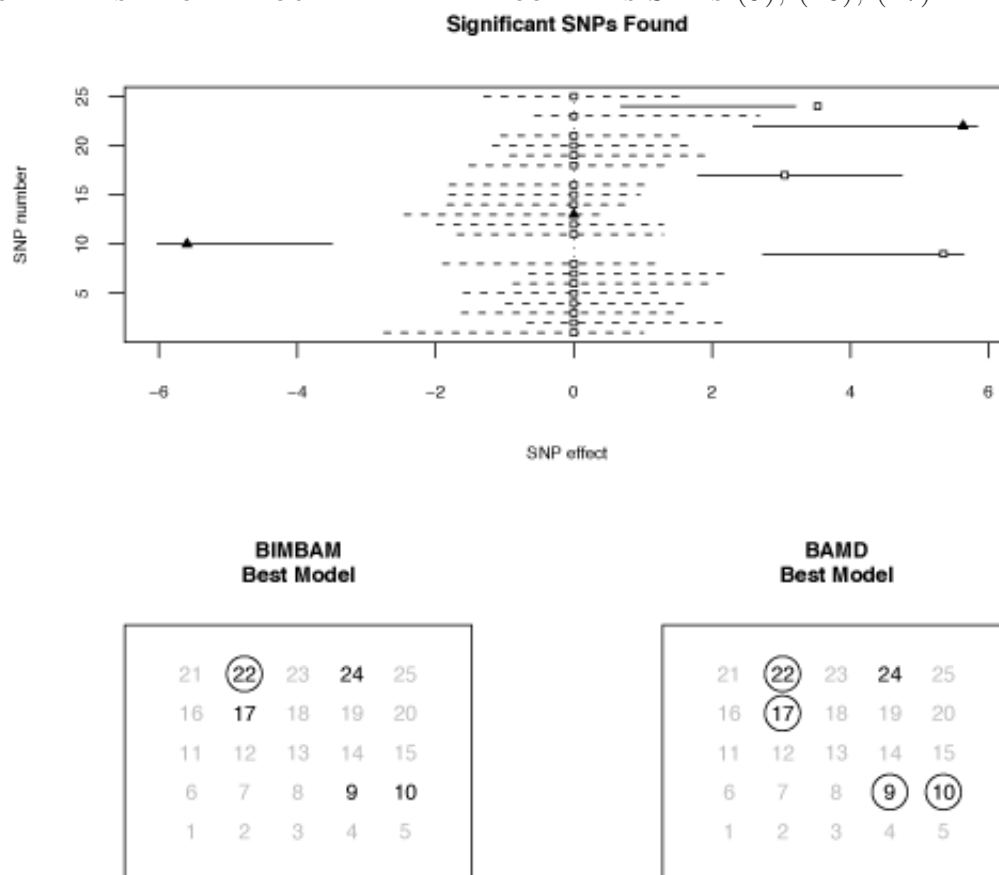
Each figure shows that BAMD significantly out performs BIMBAM. In the first example, BIMBAM found only 1 of the 14 significant SNPs, while BAMD found 6. In the second example there were 5 significant SNPs, and BIMBAM only found 1 again, while BAMD found 4 of them.

5 Analysis of the Loblolly Pine Data

Carbon isotope discrimination (CID) is a time-integrated trait measure of water use effi-
ciency. 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 ap-
proximately 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 informa-
tive priors, we infer the progeny SNP genotype through Mendelian randomization (Falconer
and Mackay 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 4 SNPs that were picked out when using the informative
prior.

We detected significant effects of several SNPs on CID at a 95% Bayesian confidence inter-
val (Table 4). Using the uninformative prior, we found 3 significant SNPs ((3) ccoaomt_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 informa-
tive prior. This suggests that for these data, the effect of the prior information is important,
but not substantial. 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 impor-

Table 4: Significant SNPs from QTDT tests and the results from the BAMD association model, with 95% confidence intervals. * 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, non-coding; UTR, untranslated region.

| SNP	Informative Prior 95% C.I.	Uninformative Prior 95% C.I.	Type‡
(3) caf1_s1*	(-0.008, 0.110)	**(0.013,0.129)**	Syn
(5) ccoaomt_s10*†	**(-0.103,-0.012)**	**(-0.097,-0.005)**	NC(intron)
(6) cpk3_s5	**(-0.052,-0.004)**	(-0.048,0.001)	Syn
(29) dhn1_s2*†	**(0.065,0.113)**	(0.044,0.092)	NC(3'UTR)
(31) ein2_s1*†	**(0.077,0.142)**	**(0.067,0.126)**	NC(3'UTR)

tant to note that BAMD detected these SNPs simultaneously, indication that their collective effect on the phenotype is being detected.

The use of tag SNPs in a pedigree does not allow for "fine mapping" to SNP effects (Neale and Ingvarsson 2008; Flint-Garcia *et al.* 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 dataset. Figures 3 is based on using uninformative SNP priors, while Figures 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 non-informative 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 4 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

Figure 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) ccoaomt, (31) ein2.
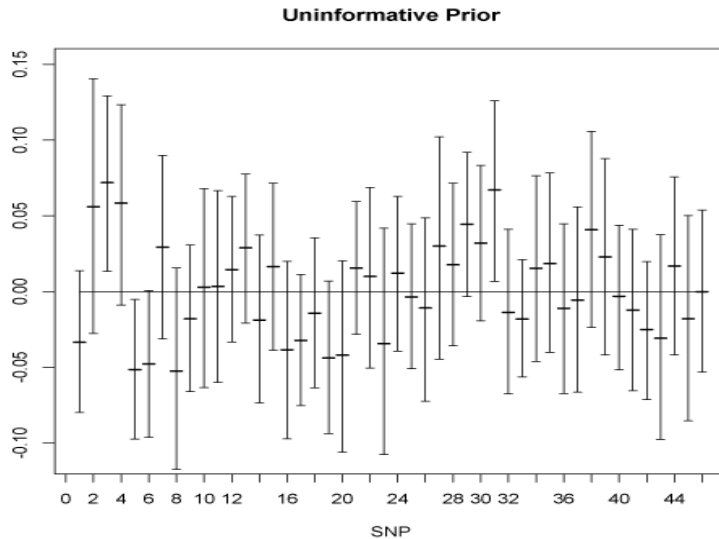

found to contain SNPs (5), (29) and (31).

6   Discussion

Association testing is being applied to discover relationships among SNPs and complex traits in plants and animals (Flint-Garcia *et al.* 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.

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.

21

Figure 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) ccoaomt, (6) cpk3, (29) dhn1, (31) ein2.
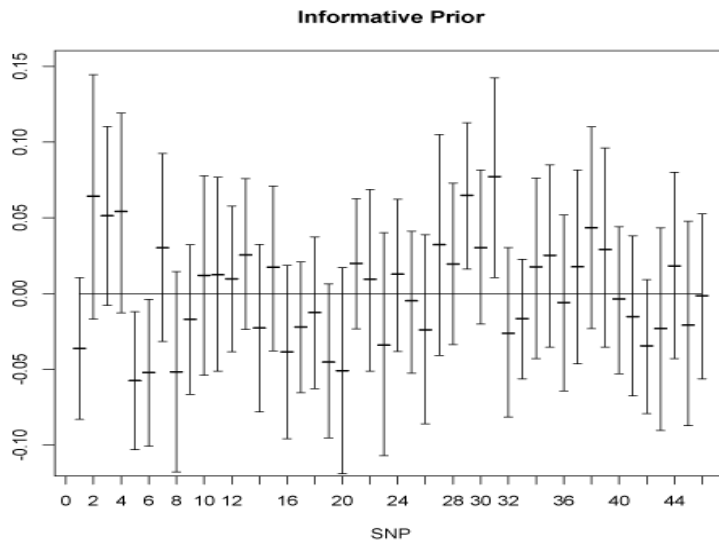
Informative Prior

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 Queseda *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 were sufficient to explain variation in the phenotype carbon isotope discrimination. This is a useful tool for biologists because 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.

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 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.

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

References

[1] Balding, D. (2006) A tutorial on statistical methods for population association studies. *Nature Reviews Genetics* **7** 781-791.

[2] Casella, G., Girón, F.J., Martínez, M.L. and Moreno E. (2009). Consistency of Bayesian procedure for variable selection. *Annals of Statistics*, **37,** 3, 1207-1228.

[3] Chen, W. M., Abecasis, G. R. (2007). Family-Based Association Tests for Genomewide Association Scan, *The American Journal of Human Genetics* **81**, 913-926.

[4] Chen, M. H. and Shao, Q. M. (1997). Estimating ratios of normalizing constants for densities with different dimensions.

[5] Dai, J. Y., Ruczinski, I., LeBlanc, M., Kooperberg, C. (2006). Imputation Methods to Improve Inference in SNP Association Studies. *Genetic Epidemiology* **30** 690-702.

[6] Falconer, D. S., and T. F. C. Macay, 1996 Introduction to Quantitative Genetics, Ed 4. Longmans Green, Harlow, Essex, UK.

[7] 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.

[8] González-Martínez SC, Ersoz E, Brown GR, Wheeler NC, Neale DB. (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.

[9] 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.

[10] Greenland S. and Finkle W. D. (1995). A critical look at methods for handling missing covariates in epidemiologic regression analyses. *Am J Epidemiol* **142** 1255-64.

[11] Hager, W. W. (1989). Updating the Inverse of a Matrix. *SIAM Review* **31** 221-239.

[12] 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.

[13] Hirschhorn, J. N. and Daly, M. J. (2005). Genome-Wide Association Studies for Common Diseases and Complex Traits. *Nature Genetics* **6** 95-108.

[14] Hobert, J., Casella, G. (996) The Effect of Improper Priors on Gibbs sampling in Hierarchical Linear Mixed Models. *Journal of the American Statistical Association* **91** 1461-1473.

[15] Huisman, M. (2000). Imputation of Missing Item Responses: Some Simple Techniques. *Quality and Quantity* **34** 331-351.

[16] Kayihan, G. C., Huber, D. A., Morse, A. M., White, T. T., Davis, J. M. (2005). Genetic Dissection of Fusiform Rust and Pitch Canker Disease Traits in Loblolly Pine. *Theory of Applied Genetics* **110**, 948 - 958.

[17] Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data.* New York: John Wiley & Sons.

[18] 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. *Nature Genetics* **39** 906-913.

[19] Meng, X. L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration . *Statistica Sinica* **6** 831-860.

[20] Neale DB, Ingvarsson PK. (2008). Population, quantitative and comparative genomics of adaptation in forest trees. *Current Opinion in Plant Biology* **11** 149-155.

[21] OHagan A. and Forster, J. (2004). *Kendalls Advanced Theory of Statistics: Vol. 2B: Bayesian Inference.* London: Arnold.

[22] Pritchard, J., K., Stephens, M., and Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics* **155:** 945-959

[23] Quaas, R. L. (1976). Computing the Diagonal Elements and Inverse of a Large Numerator Relationship Matrix, *Biometrics*, **46, 4** 949-953.

[24] 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.

[25] Quesada T, Gopal V, Cumbie WP, Eckert AJ, Wegrzyn JL, Neale DB, Goldfarb B, Huber DA, Casella G, Davis JM. 2010. Association mapping of quantitative disease resistance in a natural population of Loblolly pine (Pinus taeda L.). *Genetics* **186** 677-686.

[26] Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods, Second Edition*. New York: Springer-Verlag

[27] 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.

[28] Servin B., Stephens M (2007) Imputation-Based Analysis of Association Studies: Candidate Regions and Quantitative Traits. *PLoS Genet* **3** e114. doi:10.1371/journal.pgen.0030114

[29] Stephens, M., Smith, N. J., Donnelly, P. (2001). A New Statistical Method for Haplotype Reconstruction from Population Data, *The American Journal of Human Genetics* **68** 978-989.

[30] Su, S.Y., White, J., Balding, D. J., Coin, L. J. M. (2008). Inference of haplotypic phase and missing genotypes in polyploid organisms and variable copy number genomic regions. *BMC Bioinformatics* **9** Article Number: 513.

[31] Sun, Y.V., Kardia, S. L. R. (2008). Imputing missing genotypic data of single-nucleotide polymorphisms using neural networks. *European Journal of Human Genetics* **16** 487-495.

[32] Szatkiewicz, J. P., Beane, G. L.;, Ding, Y., Hutchins, L., de Villena, F. P., Churchill, G. A. (2008). An imputed genotype resource for the laboratory mouse. *Mammalian Genome* **19** 199-208.

[33] van der Heijden G. J., Donders A. R., Stijnen T., 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.

[34] Wear, D. N. and Greis, J. G. (2002). Southern Forest Resource Assessment: Summary of Findings. *Journal of Forestry* **100** 6-14.

[35] 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. *Annals of Applied Statistics* **4** 342 - 1364. DOI: 10.1214/09-AOAS322

[36] 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

[37] 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.

A   Calculating the Numerator Relationship Matrix

The algorithm is due to Henderson (1976) and Quaas (1977). The individuals within 61 families and the parents for the the 61 families are ordered together such that the first $1, ..., a$ subjects are unrelated and are used as a "base" population. Let the total number of subjects within families and parents of the 61 families to 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 jth and the ith subject from the above ordered subjects.

1. If both parents of the jth individual are known, say g and h, then

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

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

where R_{ji} is the cell of numerator relationship matrix in the jth row and ith column.

2. If only one parent is known for the jth subject, say it is g, then

$$R_{ji} = R_{ij} = 0.5R_{ig}, i = 1, ..., j - 1;$$

$$R_{jj} = 1.$$

3. If neither parent is known for the jth subject,

$$R_{ji} = R_{ij} = 0, i = 1, ...j - 1;$$

$$R_{jj} = 1.$$

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 the 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.

B Details of the Variable Selector Chain

B.1 Proof of Proposition 1

The proof of Proposition 1 follows from the results of Chen and Shao (1997). For completeness we give a proof here.

Proof. Recall that

$$\mathbb{E}\left[\frac{\pi_\delta(\boldsymbol{\theta}_\delta)g(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right] = \int \frac{\pi_\delta(\boldsymbol{\theta}_\delta)g(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\pi(\boldsymbol{\theta}|\boldsymbol{Y})\mathrm{d}\boldsymbol{\theta}.$$

Writing $\pi(\boldsymbol{\theta}|\boldsymbol{Y}) = p(\boldsymbol{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})/m(\boldsymbol{Y})$, $\mathrm{d}\boldsymbol{\theta} = \mathrm{d}\boldsymbol{\theta}_\delta\mathrm{d}\boldsymbol{\theta}_{\delta^c}$ and collecting terms we have

$$\begin{aligned}
\int \frac{\pi_\delta(\boldsymbol{\theta}_\delta)g(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\pi(\boldsymbol{\theta}|\boldsymbol{Y})\mathrm{d}\boldsymbol{\theta} &= \frac{1}{m(\boldsymbol{Y})}\int \pi_\delta(\boldsymbol{\theta}_\delta)\left[\int g(\boldsymbol{\theta})p(\boldsymbol{Y}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}_{\delta^c}\right]\mathrm{d}\boldsymbol{\theta}_\delta \\
&= \frac{1}{m(\boldsymbol{Y})}\int \pi_\delta(\boldsymbol{\theta}_\delta)\left[p_\delta(\boldsymbol{Y}|\boldsymbol{\theta}_\delta)\right]\mathrm{d}\boldsymbol{\theta}_\delta \\
&= \frac{m_\delta(\boldsymbol{Y})}{m(\boldsymbol{Y})} = \mathrm{BF}_\delta,
\end{aligned}$$

the Bayes factor. $\square$

B.2 Choices of the g Function

The g function of Proposition 1, is not unique, and some would work better than others in different data-sets. For example, we can replace (6) with

$$g'(\boldsymbol{\theta}) = |\boldsymbol{Q}|^{1/2}(2\pi\sigma^2)^{-d^c/2}\exp\left(-\frac{1}{2\sigma^2}\left(\frac{|\boldsymbol{\gamma}_{\delta^c}|^2}{\phi^2} + \boldsymbol{C}_\delta'\boldsymbol{Z}_{\delta^c}\boldsymbol{Q}^{-1}\boldsymbol{Z}_{\delta^c}'\boldsymbol{C}_\delta\right)\right)$$

where $\boldsymbol{Q} = (1/\phi^2)\boldsymbol{I}_{d^c} + \boldsymbol{Z}_{\delta^c}'\boldsymbol{Z}_{\delta^c}$. This choice would be desirable if if $\boldsymbol{Z}_{\delta^c}'\boldsymbol{Z}_{\delta^c}$ is not of full rank for some particular δ (which does not happen for us).

B.3 Searching for the Largest Bayes Factor

Our approach is to run the Gibbs sampler to obtain samples from the posterior distribution of $\boldsymbol{\theta}$, and then to run the Metropolis-Hastings chain described above, using the *estimated* Bayes Factors in lieu of the true Bayes Factors, which we call the empirical M-H chain. We

now show that if we run the Gibbs sampler to obtain N samples from the posterior of $\boldsymbol{\theta}$, and then run the empirical M-H chain for t steps then, as N and t go to $+\infty$, the stationary distribution of the empirical chain is $B(\delta)$.

Theorem 1. *Denote the t-step transitional kernel for the empirical M-H chain derived with N samples of the Gibbs sampler by $\hat{K}_{N,t}(\delta, \cdot)$, and let*

$$\hat{B}_N(\delta) = \frac{\widehat{BF}_\delta^N}{1 + \sum_{\delta' \neq \mathbf{1}} \widehat{BF}_{\delta'}^N}, \tag{11}$$

the probability distribution on the model space defined by the estimated Bayes factors. Then

$$||\hat{K}_{N,t}(\delta, \cdot) - B(\cdot)||_{\text{TV}} \to 0 \text{ as } N, t \to +\infty, \tag{12}$$

where $\hat{K}_{N,t}(\delta, \cdot)$ is the transition kernel of the empirical M-H chain, $B(\cdot)$ are the true Bayes factors, and $||\cdot||_{\text{TV}}$ denotes total variation norm.

Proof. Using the triangle inequality,

$$||\hat{K}_{N,t}(\delta, \cdot) - B(\cdot)|| \leq ||\hat{K}_{N,t}(\delta, \cdot) - \hat{B}_N(\cdot)|| + ||\hat{B}_N(\cdot) - B(\cdot)||, \tag{13}$$

we now show that the two terms on the right hand side can be made arbitrarily small, starting with $||\hat{K}_{N,t}(\delta, \cdot) - \hat{B}_N(\cdot)||$. We can explicitly write the transition kernel as

$$\hat{K}_{N,t}(\delta_1, \delta_2) = \hat{\rho}(\delta_1, \delta_2)q(\delta_2|\delta_1)I\{\delta_2 \neq \delta_1\} + (1 - r(\delta_1))I\{\delta_2 = \delta_1\},$$

where $\hat{\rho}(\delta_1, \delta_2) = \min\{1, \hat{BF}_{\delta_2}^N / \hat{BF}_{\delta_1}^N\}$ and $q(\delta_2|\delta_1) = q(\delta_1|\delta_2)$ by symmetry. We now show that detailed balance is satisfied with $\hat{B}_N$, that is, when $\delta_2 \neq \delta_1$,

$$\hat{\rho}(\delta_1, \delta_2)q(\delta_2|\delta_1)\hat{B}_N(\delta_1) = \hat{\rho}(\delta_2, \delta_1)q(\delta_1|\delta_2)\hat{B}_N(\delta_2), \tag{14}$$

and that when $\delta_2 = \delta_1$,

$$(1 - r(\delta_1))I\{\delta_2 = \delta_1\}\hat{B}_N(\delta_1) = (1 - r(\delta_2))I\{\delta_2 = \delta_1\}\hat{B}_N(\delta_2). \tag{15}$$

The second equality is trivial, while our symmetric candidate q leads to the following simplification in equation (14):

$$\hat{\rho}(\delta_1, \delta_2)\hat{B}_N(\delta_1) = \hat{\rho}(\delta_2, \delta_1)\hat{B}_N(\delta_2) \tag{16}$$

Suppose that $\widehat{\mathrm{BF}}_{\delta_2}^N < \widehat{\mathrm{BF}}_{\delta_1}^N$. Then,

$$\hat{\rho}(\delta_1, \delta_2)\hat{B}_N(\delta_1) = \frac{\widehat{\mathrm{BF}}_{\delta_2}^N}{\widehat{\mathrm{BF}}_{\delta_1}^N} \times \frac{\widehat{\mathrm{BF}}_{\delta_1}^N}{1 + \sum_{\delta' \neq \mathbf{1}} \widehat{\mathrm{BF}}_{\delta'}^N} = \frac{\widehat{\mathrm{BF}}_{\delta_2}^N}{1 + \sum_{\delta' \neq \mathbf{1}} \widehat{\mathrm{BF}}_{\delta'}^N} = \hat{\rho}(\delta_2, \delta_1)\hat{B}_N(\delta_2)$$

An identical argument covers the case where $\widehat{\mathrm{BF}}_{\delta_2}^N > \widehat{\mathrm{BF}}_{\delta_1}^N$. This proves detailed balance so $\hat{B}_N(\delta)$ is the stationary distribution, and the first term in (13) is becomes arbitrarily small. (This convergence is uniform, as the finite state space results in a uniformly ergodic chain.)

For the second term, by Scheffe's lemma, we know that

$$||\hat{B}_N(\cdot) - B(\cdot)|| = \frac{1}{2} \sum_{\delta \in 2^\Omega} |\hat{B}_N(\delta) - B(\delta)| \tag{17}$$

Note that the summation is a finite sum, with 2^s summands. For each δ, the ergodic theorem applied to the Gibbs sampler, tells us that as $N \to \infty$, $\widehat{\mathrm{BF}}_\delta \to \mathrm{BF}_\delta$ almost surely. Hence it follows that for each δ, $\hat{B}_N(\delta) \to B(\delta)$, as $N \to \infty$. Thus there is M such that for all $N \geq M$,

$$|\hat{B}_N(\delta) - B(\delta)| < \epsilon/2^s$$

This in turn implies, from equation (17), that

$$||\hat{B}_N(\cdot) - B(\cdot)|| = \frac{1}{2} \sum_{\delta \in 2^\Omega} |\hat{B}_N(\delta) - B(\delta)| < \epsilon/2, \tag{18}$$

we shows that we can make the total variation distance between $\hat{K}_{N,t}(\delta, \cdot)$ and $B(\cdot)$ arbitrarily small, as long as t and N are large enough. $\qquad\square$

Finally, note that from equation (7), the Bayes factor estimate requires the determinant of $\mathbf{Z}_{\delta^c}^{(i)\prime} \mathbf{Z}_{\delta^c}^{(i)}$ to be computed for each value obtained from the Gibbs sampler. This is a large computational burden, and we have replaced this with the single computation (for each δ) of a determinant of $\mathbf{Z}_{\delta^c}^\prime \mathbf{Z}_{\delta^c}$ averaged over the sampled missing values.

Theorem 2. *Let* $P_{\delta^c} := \boldsymbol{Z}_{\delta^c}(\boldsymbol{Z}_{\delta^c}'\boldsymbol{Z}_{\delta^c})^{-1}\boldsymbol{Z}_{\delta^c}'$, $\boldsymbol{C}_\delta := (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}_\delta\boldsymbol{\gamma}_\delta)$ *and*

$$g(\boldsymbol{\theta}) = (2\pi\sigma^2)^{-d^c/2}|\boldsymbol{Z}_{\delta^c}'\boldsymbol{Z}_{\delta^c}|^{1/2}\exp\left(-\frac{1}{2\sigma^2}\boldsymbol{C}_\delta'P_{\delta^c}\boldsymbol{C}_\delta\right) \tag{19}$$

Then if $\boldsymbol{\theta}^{(i)}$ *are samples from the posterior* $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$, *the estimator*

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

is strongly consistent for BF_δ. *The samples* $\boldsymbol{\theta}^{(i)}$ *could be independent, or from an ergodic Markov chain with stationary distribution* $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$.

Proof. It is straightforward to show that g satisfies the requirements of Proposition 1. Then depending on whether the samples $\boldsymbol{\theta}^{(i)}$ are independent or from a Markov Chain, we can apply the Strong Law of Large Numbers or the Ergodic Theorem together with Proposition 1 to obtain the result. $\qquad\square$

C   Convergence of the Gibbs Sampler

Suppose we are running a Gibbs sampler in which one cycle consists of updating (X, Y, Z), and the $Z_{n\times p}$ matrix can be decomposed as $(Z_1, Z_2, \ldots, Z_p)$. Standard Gibbs sampling updates $Z_{n\times p}$ all together in one cycle, in contrast, we can just systematically update one column of Z at each cycle. By updating one column of Z in each iteration, the computation time can decrease dramatically, especially when the number of columns of the Z matrix is in the order of hundreds or thousands. In this appendix we show that this updating scheme preserves the ergodicity of the original sampler.

Here is the updating scheme. Start with value

$$(X^{(0)}, Y^{(0)}, Z_1^{(0)}, ..., Z_p^{(0)}).$$

In the first cycle, conditional on

$$(Y^{(0)}, Z_1^{(0)}, ..., Z_p^{(0)}) \quad \text{update } X^{(0)} \text{ and get} \quad (X^{(1)}, Y^{(0)}, Z_1^{(0)}, ..., Z_p^{(0)}).$$

Then conditional on

$$(X^{(1)}, Z_1^{(0)}, ..., Z_p^{(0)}), \quad \text{update } Y^{(0)} \text{ and get} \quad (X^{(1)}, Y^{(1)}, Z_1^{(0)}, ..., Z_p^{(0)}).$$

Finally, conditional on

$$(X^{(1)}, Y^{(1)}, Z_2^{(0)}, ..., Z_p^{(0)}), \quad \text{update } Z_1^{(0)} \text{ and get} \quad (X^{(1)}, Y^{(1)}, Z_1^{(1)}, Z_2^{(0)}, ..., Z_p^{(0)}).$$

The above is one iteration of the Gibbs sampler. In the same manner we will update $(X^{(1)}, Y^{(1)}, Z_2^{(0)})$ in the second cycle. Continuing in this manner, after the *pth* cycle, we will get

$$(X^{(p)}, Y^{(p)}, Z_1^{(1)}, Z_2^{(1)}, ..., Z_p^{(1)}),$$

and thus one complete iteration of the Markov chain is

$$(X^{(0)}, Y^{(0)}, Z_1^{(0)}, ..., Z_p^{(0)}) \to (X^{(p)}, Y^{(p)}, Z_1^{(1)}, Z_2^{(1)}, ..., Z_p^{(1)}).$$

To prove that the above updating scheme conserves the ergodicity property, we will show first that the kernel of the Gibbs sampler preserves the original stationary distribution condition and, furthermore, we will show that with any initial condition the chain converges to the target stationary distribution.

Let $K((x^{(0)}, y^{(0)}, z_0^{(0)}, ..., z_p^{(0)}), (x^{(p)}, y^{(p)}, z_1^{(1)}, ..., z_p^{(1)}))$ denote the transition kernel of the sampler. To prove that the Gibbs sampler has the correct stationary distribution, we need to show that the following equation holds:

$$
\begin{aligned}
&f(x^{(p)}, y^{(p)}, z_1^{(1)}, \ldots, z_p^{(1)}) \\
&= \int K((x^{(0)}, y^{(0)}, z_0^{(0)}, \ldots, z_p^{(0)}), (x^{(p)}, y^{(p)}, z_1^{(1)}, ..., z_p^{(1)})) \\
&\qquad \times f(x^{(0)}, y^{(0)}, z_1^{(0)}, \ldots, z_p^{(0)}) \; dx^{(0)} dy^{(0)} dz_1^{(0)} \cdots dz_p^{(0)},
\end{aligned}
\tag{20}
$$

where f is the stationary distribution of the original chain. As K consists of the successive conditional distributions, the integrations are straightforward and (20) holds. Thus, updating one column per cycle preserves the original stationary condition. Ergodicity of the chain

follows from Theorem 10.10 in Robert and Casella (2004). Finally, we note that the subvector $(X^{(t)}, Y^{(t)})$ converges to its own marginal distribution, which is what we are interested in. So if we are just interested in the estimates of X, Y only, it is legitimate to use samples $(X^{(t)}, Y^{(t)}), t = 1, 2, \cdots$, instead of $(X^{(pt)}, Y^{(pt)}), t = 1, 2, \cdots$.

Applying this argument to the Gibbs sampler of Section 2.2, we have proved the following result.

Theorem 1. *Let Z_j denote the jth column in the $Z_{n \times p}$ matrix for SNPs (the jth SNP for all observations). For the Gibbs sampler corresponding to (2) and (3), if instead of updating all parameters $\left(\beta^{(t)}, \gamma^{(t)}, Z_{n \times p}^{(t)}, \sigma^{2(t)}, \phi^{2(t)} \right)$ in each cycle, we just update $\left(\beta^{(t)}, \gamma^{(t)}, Z_j^{(t)}, \sigma^{2(t)}, \phi^{2(t)} \right)$ in each iteration, the Markov Chain converges to the same target stationary distribution and ergodicity also holds.*

D   Matrix Inverse Updates

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

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

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

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

$$A_k^{-1} = A_{k-1}^{-1} - \frac{A_{k-1}^{-1} u_k v_k' A_{k-1}^{-1}}{1 + v_k' A_{k-1}^{-1} u_k}. \tag{21}$$

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 multiplications of matrices by vectors. Moreover, in many applications the vectors u_k, v_k are sparse, so the multiplications amount to extracting elements.

E   Estimation with the EM Algorithm

E.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

$$L_C \quad \propto \quad \prod_{i \in I_0} \exp\left(-\frac{1}{2\sigma^2}(Y_i - X_i\beta - Z_i\gamma)^2\right) \prod_{i \in I_M} \exp\left(-\frac{1}{2\sigma^2}(Y_i - X_i\beta - Z_i^*\gamma)^2\right) \quad (22)$$

where I_o indexes those individual 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

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

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

$$P(Z_i^*) = \frac{\exp\left(-\frac{1}{2\sigma^2}(Y_i - X_i\beta - Z_i^*\gamma)^2\right)}{\sum_{Z_i^*} \exp\left(-\frac{1}{2\sigma^2}(Y_i - X_i\beta - Z_i^*\gamma)^2\right)}, \quad (23)$$

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 (label at left of table)

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).

E.2 An EM Algorithm

To the expected value of the log of the complete data likelihood (22), 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

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

where

$$\mathrm{E}(Z_i^*) = (Z_i^o, \mathrm{E}(Z_i^m)) \text{ and } \mathrm{Var}(Z_i^*\gamma) = \mathrm{Var}(Z_i^m\gamma_i^m). \tag{24}$$

If we define

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

the expected complete data log likelihood is

$$\mathrm{E}\log L_C = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}|Y - X\beta - Z\gamma|^2 - \frac{1}{2\sigma^2}\gamma' V_Z\gamma \tag{25}$$

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

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

and $V_Z = \sum_{i\in I_M} V_{Z_i}$. Standard calculus will show that the MLEs from (25) are given by

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'(Y - Z\hat{\gamma}) \\ \hat{\gamma} &= (Z'Z - V_Z)^{-1}Z'(I - H)Y \\ \hat{\sigma}^2 &= \frac{1}{n}\left(|Y - X\hat{\beta} - Z\hat{\gamma}|^2 + \hat{\gamma}'V_Z\hat{\gamma}\right). \end{aligned} \tag{26}$$

The algorithm now iterates between (23), (24), and (26) until convergence.

E.3   Implementation

To implement the EM algorithm we must be able to either

1. Calculate the expectation and variance in (24), or

2. Generate a random sample from (23) and calculate the terms in (24) by simulation.

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

Going back to (23), 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 $Z_i^m = (Z_{i1}^m, \ldots, Z_{ik}^m)$, and for $\mathbf{c} = (c_1, \ldots, c_k)$,

$$P(Z_i^m = \mathbf{c}_0) = \frac{\exp\left(-\frac{1}{2\sigma^2}(Y_i - X_i\beta - Z_i^o\gamma_i^o - \mathbf{c}_0\gamma_i^m)^2\right)}{\sum_{\text{all } \mathbf{c}} \exp\left(-\frac{1}{2\sigma^2}(Y_i - X_i\beta - Z_i^o\gamma_i^o - \mathbf{c}\gamma_i^m)^2\right)},$$

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 sampler (2) and (3) needs 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.

F Supplemental Information: Comparing BAMD to BIMBAM

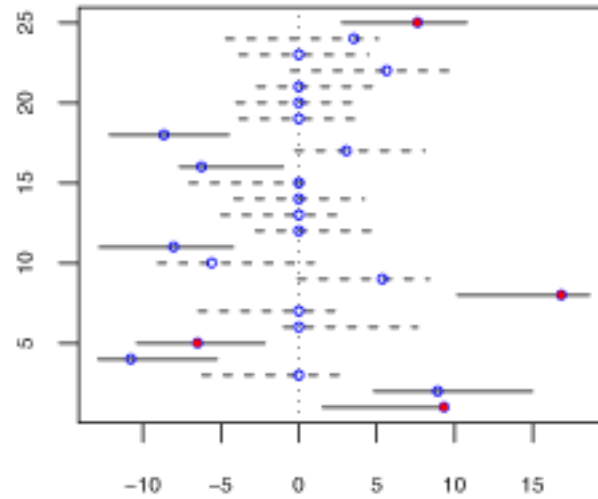## F.1 Performance under differing degrees of missing data

Once the $\boldsymbol{Z}$ matrix was generated, 10%, 20% and 30% of the values were set to be missing. The true SNPs for this experiment were $1, 2, 4, 5, 8, 9, 10, 11, 16, 17, 18, 22, 24, 25$.

F.1.1 10% missing data

Significant SNPs Found

Significant SNPs Found

BIMBAM Best Model

BAMD Best Model

F.1.3   30% missing data

Significant SNPs Found
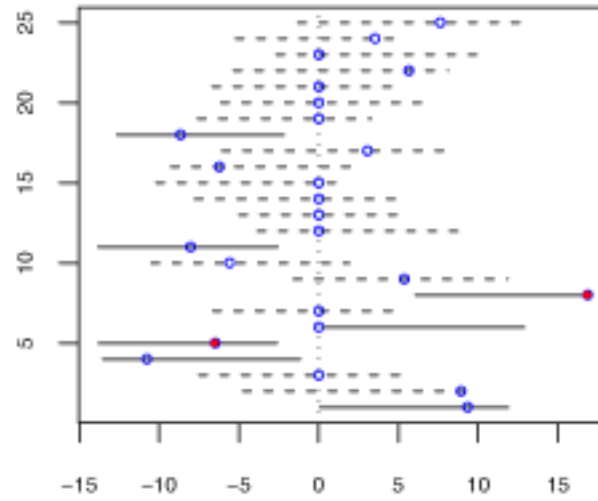
BIMBAM
Best Model

BAMD
Best Model

F.2 Performance in relation to varying magnitudes of γ effects

This portion of the experiment seeks to test how sensitive the procedures are to finding significant effects. For the first experiment any generated γ's that had absolute values less than one were set to zero. For the second experiment, the cut-off value was set to be 5. When testing other factors, the cut-off was fixed at 3.
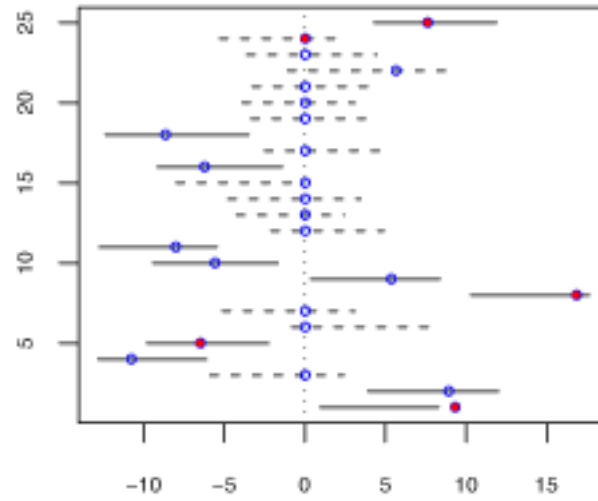
For Section F.2.1, the true SNPs were $1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19,$ $20, 22, 24, 25$. For Section F.2.2, the true SNPs were $1, 2, 4, 5, 8, 9, 10, 11, 16, 18, 22, 25$.

F.2.1 γ cut-off set at 1

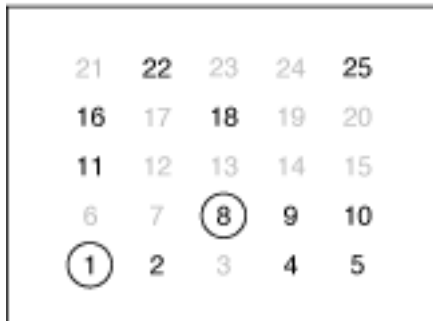**Significant SNPs Found**

BIMBAM
Best Model
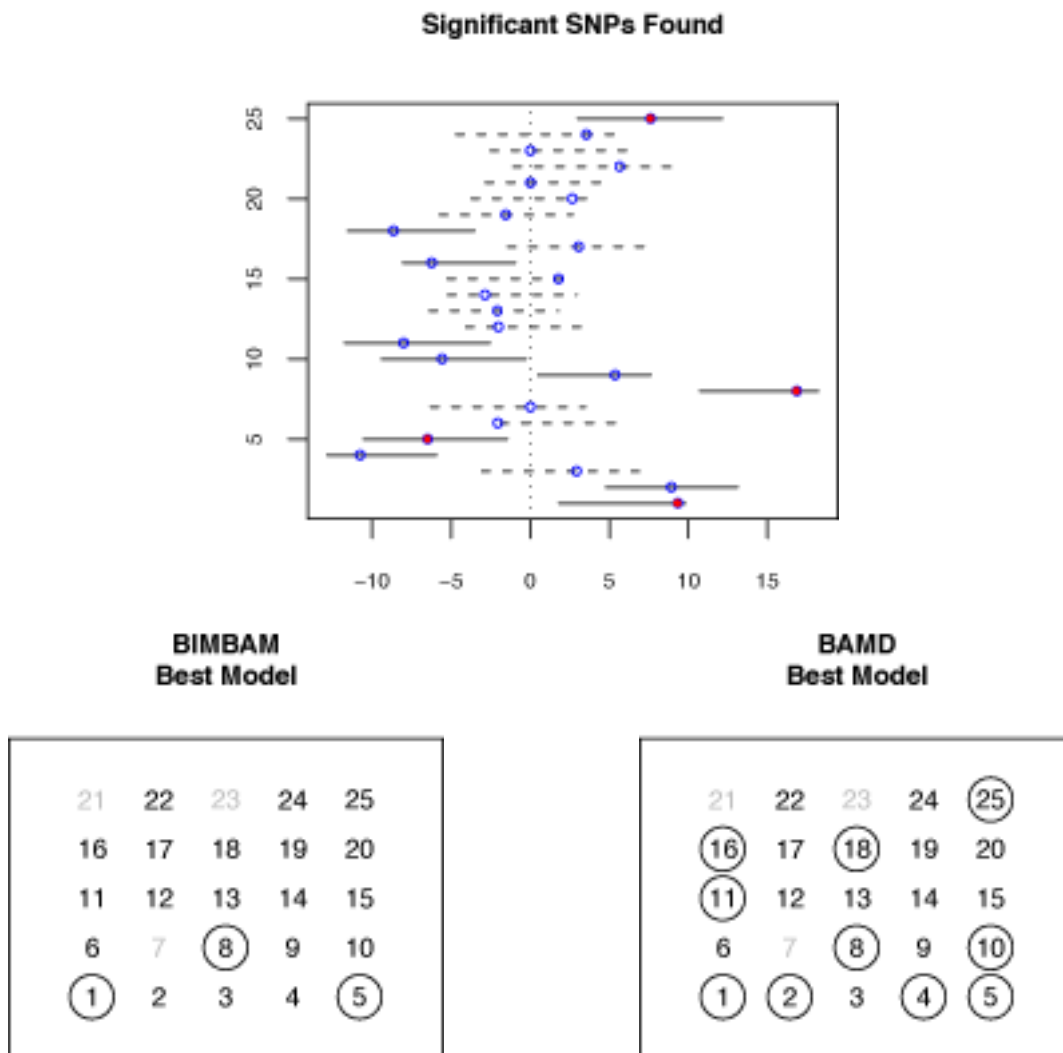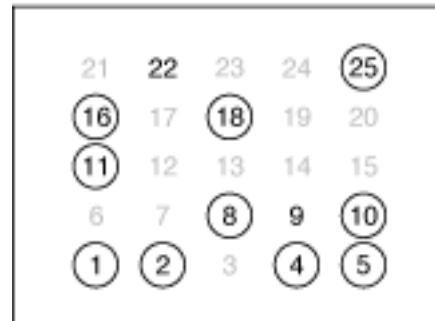
BAMD
Best Model

F.2.2 γ cut-off set at 5

Significant SNPs Found

BIMBAM
Best Model

BAMD
Best Model

F.3 Comparing performance with differing magnitudes of correlation within the family

For this part, an equicorrelation structure was assumed within each of the 3 families. The correlations used were $\rho = 0.2$ and $\rho = 0.8$. When testing other factors, all ϵ's were assumed to be independent. Again, the true SNPs were set to be $1, 2, 4, 5, 8, 9, 10, 11, 16, 17, 18, 22, 24, 25$.

Significant SNPs Found

BIMBAM
Best Model

BAMD
Best Model