Armero, C., Cabras, S., Castellanos, M. E., & Quirós, A. (2019). Two-Stage Bayesian Approach for GWAS With Known Genealogy. *Journal of Computational and Graphical Statistics*, 28 (1), pp. 197-204

DOI: 10.1080/10618600.2018.1483828

Two-Stage Bayesian Approach for GWAS With Known Genealogy

Carmen Armero [a], Stefano Cabras [b,c], María Eugenia Castellanos [d,e], and Alicia Quirós [f]

[a]Department of Statistics and Operations Research, Universitat de València, València, Spain; [b]Department of Statistics, Universidad Carlos III de Madrid, Madrid, Spain; [c]Department of Mathematics and Informatics, Università degli Studi di Cagliari, Cagliari CA, Italy; [d]Department of Informatics and Statistics, Universidad Rey Juan Carlos, Madrid, Spain; [e]Department of Economic Science, Università degli Studi di Cagliari, Cagliari CA, Italy; [f]Department of Mathematics, Universidad de León, León, Spain

ABSTRACT

Genome-wide association studies (GWAS) aim to assess relationships between single nucleotide polymorphisms (SNPs) and diseases. They are one of the most popular problems in genetics, and have some peculiarities given the large number of SNPs compared to the number of subjects in the study. Individuals might not be independent, especially in animal breeding studies or genetic diseases in isolated populations with highly inbred individuals. We propose a family-based GWAS model in a two-stage approach comprising a dimension reduction and a subsequent model selection. The first stage, in which the genetic relatedness between the subjects is taken into account, selects the promising SNPs. The second stage uses Bayes factors for comparison among all candidate models and a random search strategy for exploring the space of all the regression models in a fully Bayesian approach. A simulation study shows that our approach is superior to Bayesian lasso for model selection in this setting. We also illustrate its performance in a study on Beta-thalassemia disorder in an isolated population from Sardinia. Supplementary Material describing the implementation of the method proposed in this article is available online.

KEYWORDS
Bayes factor; Beta-thalassemia disorder; Gaussian Markov random field; Kinship coefficient; Model selection; Robust prior distribution

1. Introduction

Genome-wide association studies (GWAS) collect data on single nucleotide polymorphisms (SNPs)—genetic markers—across the genome with the aim to identify causal variants related to diseases (Balding 2006; Wagner 2013). These relationships are statistically represented by regression models which usually incorporate a large number of SNPs. Additional issues arise when individuals from the observed cohort are linked by family ties, the framework in which we focus on.

The number of SNPs is generally much larger than the number of subjects studied ($p \gg n$), which is known as an ill-posed problem. These problems originated in mathematical settings and are well known in the scientific literature. Once disseminated to the statistical world, they provided many different approaches and concepts (Wahba 1990; Girosi, Jones, and Poggio 1993; Nychka 2000) which generally aim to reconstruct a whole function from noisy observations. One of the most popular approaches to the subject is regularization (O'Sullivan 1986; Eilers and Marx 1996; Ruppert, Wand, and Carroll 2003) which relies on controlling overfitting through a roughness penalty. The lasso, introduced by Tibshirani (1996), is a widely used shrinkage method for linear regression models which minimizes the sum of squared errors with a smoothing parameter $\lambda > 0$ on a penalty defined as the sum of absolute values of the regression coefficients. Lasso shrinks some coefficients but also sets others to zero, thus providing a subset of predictors that are the outcome of the lasso model selection procedure. There are other proposals to penalize the likelihood in these large p problems such as ridge regression (Hoerl and Kennard 1988), bridge regression (Frank and Friedman 1993; Fu 1998), the elastic net regularization method (Zou and Hastie 2005), etc. Most of these methods have been implemented in well-used software as MERLIN (Abecasis et al. 2002), glmnet (Friedman, Hastie, and Tibshirani 2010), and many others cited in applied and review studies in genetics that are not focused on the fundamentals of the statistical methods developed herein. Some examples of these can be found in Benyamin, Visscher, and McRae (2009), Ott, Kamatani, and Lathrop (2011), and Herold et al. (2016).

Bayesian reasoning accounts for penalization through the prior distribution for the parameters of the model. Several articles discuss and analyze the Bayesian version of the regularization methods mentioned above, as the Bayesian lasso introduced in Park and Casella (2008) or the Bayesian elastic net in Li et al. (2010). Other Bayesian solutions to this problem propose a different type of sparse priors, for example, the spike and slab prior (George and McCulloch 1993; Ishwaran and Rao 2005) or the general class of sparse priors proposed in Castillo, Schmidt-Hieber, and van der Vaart (2015), and references therein. However, Castillo, Schmidt-Hieber, and van der Vaart (2015) also pointed out that the lasso regularization is essentially non-Bayesian in the sense that the corresponding full posterior distribution is useless for uncertainty quantification. For this reason, we have avoided this technique and propose, in this article, a fully Bayesian second stage after a first stage of dimension reduction.

GWAS can be viewed as a model selection problem. The procedure for model comparison within the Bayesian reasoning is the Bayes Factor (BF) (Kass and Raftery 1995). As a part of the framework of linear regression models, results using BF are very sensitive to the specified prior distribution over model parameters, especially to those parameters that are not common to all the models, such as the regression coefficients. This property was studied by Kass and Raftery (1995) and Berger and Pericchi (2001), showing that the above-mentioned sensitivity does not vanish as the sample size grows. Furthermore, improper prior distributions, frequently used in estimation theory, are invalidated for BF and the use of "arbitrary" proper vague priors is not advisable for model selection (see sec. 1.5 in Berger and Pericchi 2001). Bayarri et al. (2012) explored this question and proposed a desideratum of properties that prior distributions over parameters must verify for model comparison. In addition, they also propose the robust prior distribution that verifies these properties for model comparison in the linear regression model. Besides, when considering a large number of explanatory variables, multiplicity issues can be accounted for by choosing an adequate prior over the model space, like the hierarchical prior, proposed by Scott and Berger (2010). Moreover, in GWAS, enumerating all possible models becomes cumbersome due to the size of the model space. García-Donato and Martínez-Beneito (2013) reviewed some of the strategies proposed in the literature, showing that the empirical search strategy based on Gibbs sampling (George and McCulloch 1997) produces the best results.

Yazdani and Dunson (2015) proposed a multi-stage design to manage the intractability of variable selection in GWAS, by accounting, at the same time, for family relationships in the sample. Family-based GWAS deal with studies where individuals are linked by kinship ties. A clear example of this occurs in animal breeding studies or genetic diseases in isolated populations with highly inbred individuals. As the usual linear regression model assumes independence between subjects, random effects can be added to the model to connect the related individuals and assess the relevance of the latent elements in the variability of the data. The most commonly used measure of relatedness between two individuals is the kinship coefficient, which is defined as the probability that two genes sampled at random from each individual are identical (Malecot 1948).

We propose a family-based two-stage GWAS. In the first stage, the genetic relatedness between individuals is taken into account to reduce the dimension of the problem by selecting promising SNPs through individual regression analyses. This selection procedure is based on credible intervals and not BF because the inferential processes are based on improper objective prior distributions. We refine the SNP selection in the second stage. We propose a fully Bayesian regression model and BF for model selection with a random search strategy for exploring the space of all models. Our reasons for working with a fully Bayesian alternative to lasso in the second stage are two-fold. On the one hand, the lasso is essentially non-Bayesian as mentioned above. On the other hand, the fact that the lasso uses an identical penalization on each regression coefficient can produce bias in the resulting estimates (Lee et al. 2012).

This article is organized as follows. Section 2 contains the proposed statistical model and details about its inferential process. Section 3 discusses the two-stage model on a study about the beta-thalassemia—an inherited blood disorder—a toy example to better illustrate the two-stage proposal, and a simulation study. This section also includes a comparison of our approach with the Bayesian lasso and GEMMA software (Zhou, Carbonetto, and Stephens 2013), and an evaluation of the inclusion of a family effect. Conclusions are presented in Section 4.

2. The Statistical Model

Let $\boldsymbol{y} = (y_1, \ldots, y_n)'$ be the vector of values of the response variable representing the amount of disease for a sample of n individuals. A GWAS for n related individuals can be expressed as a linear mixed model that describes $\boldsymbol{y}$ as a function of p SNPs and some measurement on the familiar dependencies among the sampled individuals in the form

$$\boldsymbol{y} = \mathbf{1}\beta_0 + \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{W} + \boldsymbol{\epsilon}, \tag{1}$$

where $\tilde{\boldsymbol{X}}$ is the $(n \times p)$ matrix containing the information in the p SNPs for all subjects, $\boldsymbol{\beta}$ is the vector of the unknown regression coefficients associated with the SNPs, $\boldsymbol{W}$ is an $(n \times 1)$ random effect vector which describes the family relations among the individuals in the sample, and $\boldsymbol{\epsilon}$ is a normally distributed vector of measurement error. Information provided by the three possible values for each SNP, $\{aa, aA, AA\}$, is encoded in $\tilde{\boldsymbol{X}}$ as the number of A's. The dimension of the typical GWAS makes it impossible to estimate this full model. So we proceed in two stages.

2.1. First Stage: Dimension Reduction

We will study the association of the disease with each SNP separately taking into account the family relationship of the individuals in the sample. Thus, for the jth SNP, we consider the regression model

$$\boldsymbol{y} = \mathbf{1}\beta_0^{(j)} + \tilde{\boldsymbol{X}}^{(j)}\beta^{(j)} + \boldsymbol{W}^{(j)} + \boldsymbol{\epsilon}^{(j)}, \tag{2}$$

where $\tilde{\boldsymbol{X}}^{(j)}$ is now the vector that only includes information about the value of the jth SNP for each individual in the sample (i.e., the jth column of the $\tilde{\boldsymbol{X}}$ matrix in (1)) with unknown regression coefficient parameter $\beta^{(j)}$ and $\boldsymbol{\epsilon}^{(j)} \sim \mathcal{N}(\mathbf{0}, \sigma_j^2\boldsymbol{I})$ is an $(n \times 1)$ vector of random errors. Here vector $\boldsymbol{W}^{(j)}$ is also a random effects vector for modeling the family relations of the individuals in the sample. We chose it as a Gaussian Markov random field with mean $\mathbf{0}$ and covariance matrix $\sigma_{wj}^2 K$ which contains a general element variance, σ_{wj}^2, and a matrix K accounting for the kinship coefficient between all the pairs of individuals in the sample.

The kinship coefficient is the simplest measure of the relationship between two relatives. It varies between 0 and 1/2. The kinship coefficient is 0 for unrelated individuals, 1/2 for individuals with themselves, 1/4 between parent and child, 1/8 between aunt/uncle and nephew/niece and grandparents and grandchildren, etc. Figure 1 represents the kinship matrix for individuals in the toy example, in which the family structure is depicted in Figure 3. The kinship matrix defines a neighborhood structure in the population studied that can be naturally incorporated into the model. In fact, with this variance-covariance

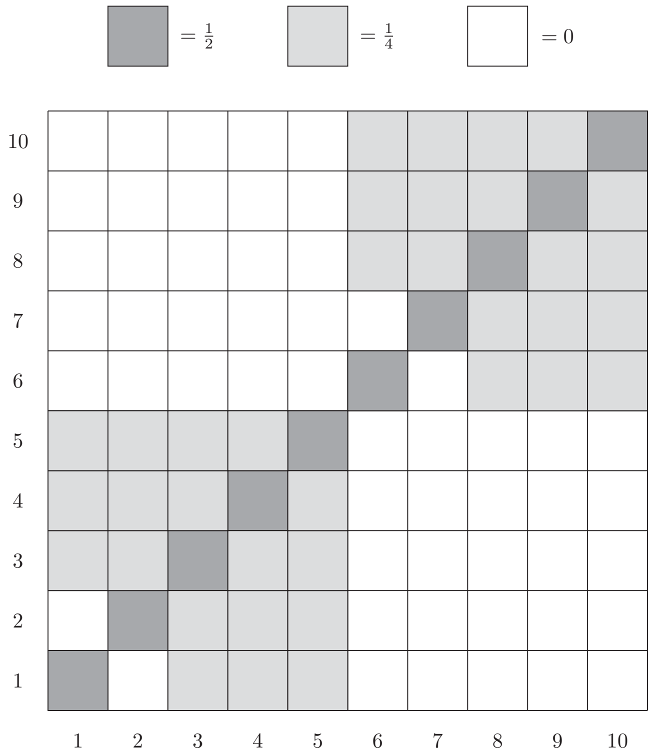

Figure 1. Kinship matrix corresponding to individuals in the toy example depicted in Figure 3. In this example, the main pattern is a two block structure, that individuals fall into two kinship groups.

matrix we can assume that the variability of the random effects associated with related individuals is greater than unrelated individuals, hoping to reduce the possible confounding effect on the response and accounting for the dependence between the subjects in the sample.

We elicit a prior distribution for the parameters and hyperparameters of the model to complete the Bayesian model. We assume a prior independence default scenario with marginal objective prior distributions

$$\begin{aligned} \pi(\beta^{(j)}) &\propto 1, \\ \pi(\sigma_j^2) &\propto \frac{1}{\sigma_j^2}, \\ \pi(\sigma_{wj}^2) &\propto \frac{1}{\sigma_{wj}^2}, \ j = 1, \dots, p. \end{aligned} \tag{3}$$

The improper condition of these prior distributions makes BF not adequate for SNP selection in this case (see sec. 1.5 in Berger and Pericchi 2001). We use `INLA` (Rue, Martino, and Chopin 2009)—the `R-INLA` package (*www.r-inla.org*)—to make inference about the unknown quantities of the model. The 95% credible intervals for the regression coefficients corresponding to each SNP are used to select the promising SNPs. Only SNPs whose interval does not contain the 0 will be included in the second stage.

2.2. Second Stage: Model Selection

We approach model selection considering all possible regression models constructed through all the 2^{p_s} subsets of the set of p_s selected SNPs in the first stage. We define a latent random vector of binary variables $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_{p_s})'$, where $\gamma_j = 1$ indicates that SNP j, $j = 1, \dots, p_s$, is present in model $M_{\boldsymbol{\gamma}}$, and $\gamma_j = 0$ otherwise. For each $\boldsymbol{\gamma}$, we consider $k_{\boldsymbol{\gamma}} = \sum \gamma_j$ as the number of SNPs in model $M_{\boldsymbol{\gamma}}$, and $\boldsymbol{X}_{\boldsymbol{\gamma}}$ as the design matrix corresponding to model $M_{\boldsymbol{\gamma}}$, which is more precisely defined as

$$M_{\boldsymbol{\gamma}} : \boldsymbol{y} = \boldsymbol{1}\beta_0 + \boldsymbol{X}_{\gamma}\boldsymbol{\beta}_{\gamma} + \boldsymbol{\epsilon}, \ \ \boldsymbol{\gamma} \in \{0, 1\}^{p_s}, \tag{4}$$

with $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$. In addition, we define the null model, M_0, as $\boldsymbol{y} = \boldsymbol{1}\beta_0 + \boldsymbol{\epsilon}$. Parameters (β_0, σ) are common to all models, while $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ are model specific.

Under model M_0 the prior distribution for $(\beta_0, \sigma)'$ is $\pi(\beta_0, \sigma)$. We express the prior distribution for $(\beta_0, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma)'$ under model $M_{\boldsymbol{\gamma}}$ as

$$\pi_{\boldsymbol{\gamma}}(\beta_0, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma) = \pi_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \beta_0, \sigma)\, \pi(\beta_0, \sigma).$$

We use the "robust prior distribution" proposed in Bayarri et al. (2012) based on the group invariance criterion and predictive matching criterion. It specifies improper priors over the common intercept and standard deviation, $\pi(\beta_0, \sigma) = 1/\sigma$, and robust priors for the conditional prior distribution $\pi_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \beta_0, \sigma)$ that cannot be improper or vague to obtain appropriate BF (Berger and Pericchi 2001). Especially

$$\pi_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \beta_0, \sigma) = \int_0^\infty \mathcal{N}(\boldsymbol{\beta}_{\boldsymbol{\gamma}} | 0, g\, \Sigma_{\boldsymbol{\gamma}}) f_{\boldsymbol{\gamma}}(g) dg, \tag{5}$$

where $\Sigma_{\boldsymbol{\gamma}} = \mathrm{cov}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}) = \sigma^2 (\boldsymbol{X}_{\boldsymbol{\gamma}}^t (\boldsymbol{I} - n^{-1} \boldsymbol{1}\boldsymbol{1}^t) \boldsymbol{X}_{\boldsymbol{\gamma}})^{-1}$ is the variance-covariance matrix of the maximum likelihood estimator of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, and

$$f_{\boldsymbol{\gamma}}(g) = \frac{1}{2} \left(\frac{n+1}{k_{\boldsymbol{\gamma}} + 1} \right)^{1/2} (g+1)^{-3/2} 1_{\{g > \frac{n+1}{k_\gamma + 1} - 1\}}.$$

The posterior probability for each model $M_{\boldsymbol{\gamma}}$, can be expressed as

$$P(M_{\boldsymbol{\gamma}} \mid \mathcal{D}) = \frac{B_{\boldsymbol{\gamma}0}}{1 + \sum_{\boldsymbol{\gamma}'} B_{\boldsymbol{\gamma}'0} P_{\boldsymbol{\gamma}'0}}, \tag{6}$$

where $\mathcal{D}$ represent the data, $P_{\boldsymbol{\gamma}0}$ is the prior odds $P_{\boldsymbol{\gamma}0} = P(M_{\boldsymbol{\gamma}})/P(M_0)$, and $B_{\boldsymbol{\gamma}0}$ is the Bayes factor of model $M_{\boldsymbol{\gamma}}$ to M_0. We adopt the proposal by Scott and Berger (2010),

$$P_{\boldsymbol{\gamma}0} = \binom{p_s}{k_{\boldsymbol{\gamma}}}^{-1},$$

for the selection of these prior odds. One of the advantages of the election of the "robust prior" is that it provides closed-form expressions for BFs, which is suitable for the analysis of a large number of models. In particular, the BF of $M_{\boldsymbol{\gamma}}$ to M_0 is

$$B_{\boldsymbol{\gamma}0} = \frac{1}{k_{\boldsymbol{\gamma}} + 1} \left(\frac{n+1}{k_{\boldsymbol{\gamma}} + k_0} \right)^{-k_\gamma/2} Q_{\boldsymbol{\gamma}_0}^{-(n-k_0)/2}\, \mathrm{SH}_{\boldsymbol{\gamma}}, \tag{7}$$

where $Q_{\boldsymbol{\gamma}_0} = \mathrm{SSE}_{\boldsymbol{\gamma}}/\mathrm{SSE}_0$ is the ratio of the sum of squared errors of models $M_{\boldsymbol{\gamma}}$ and M_0, and $SH_{\boldsymbol{\gamma}}$ is the standard hypergeometric function (Gradshteyn and Ryzhi 1965)

$$\mathrm{SH}_{\boldsymbol{\gamma}} = {}_2F_1 \left(\frac{k_{\boldsymbol{\gamma}} + 1}{2}; \frac{n - k_0}{2}; \frac{k_{\boldsymbol{\gamma}} + 3}{2}; \frac{(1 - Q_{\boldsymbol{\gamma}_0}^{-1})(k_{\boldsymbol{\gamma}} + k_0)}{n+1} \right).$$

Even after the dimension reduction stage, if the number of possible models when exploring p_s SNPs is very large (2^{p_s}), it will be practically impossible to enumerate all possible models and compute all the relevant BFs. For this reason, we adopt an empirical search strategy for exploring the model space that avoids the problem of computing the posterior probability associated with each of the 2^{p_s} models (George and McCulloch 1997). This procedure uses the Gibbs sampler to generate a sample from the posterior distribution $\pi(\boldsymbol{\gamma} \mid \mathcal{D})$. In particular, we consider the sampling scheme proposed by García-Donato and Martínez-Beneito (2013), which takes advantage of the expression of the BF in (7) to obtain a sample of models which converges to $P(M_{\boldsymbol{\gamma}} \mid \mathcal{D})$. This method has been implemented using the R library `BayesVarSel`" (Garcia-Donato and Forte 2017). Throughout the article we will refer to it as Bayesian Variable Selection (BVS) method.

The vector $\boldsymbol{\gamma}$ that maximizes $P(M_{\boldsymbol{\gamma}} \mid \mathcal{D})$ leads to the highest posterior probability model, that is, the most probable according to data. There are other quantities of interest than can provide not only a complementary vision of the problem but they could also play a major role in the final SNP selection. Such are the cases of the inclusion probabilities and the median probability model. For a given explanatory variable, the inclusion probability is defined as the $\sum P(M_{\boldsymbol{\gamma}} \mid \mathcal{D})$ for all the models that contain that covariate. This is a very useful probability when the number of models is large and the posterior probability associated with the different models is very small. The median probability model is the model having covariates with inclusion probability greater than 0.5 (Barbieri and Berger 2004). (See the supplementary material for a comprehensive description of the method's implementation.)

The method proposed here assumes that $p_s < n$ after the dimension reduction stage, as it is not defined when the final number of selected SNPs is greater than the number of individuals in the sample. In this case, we recommend using a different prior distribution on the odds for each model that in a certain way prevents the Gibbs sampler from visiting models, $M_{\boldsymbol{\gamma}}$, where $k_{\boldsymbol{\gamma}} > n$. For example, Shin, Bhattacharya, and Johnson (2018) studied the performance of nonlocal priors for variable selection in $p \gg n$ settings by reducing the search space to those models with $k_{\boldsymbol{\gamma}}$ covariates, where $k_{\boldsymbol{\gamma}} < n$.

3. Results

3.1. Beta-Thalassemia Data

Beta-thalassemia is a genetic disorder caused by a mutation inside the beta-hemoglobin gene (Trecartin et al. 1981). Only homozygous individuals for the mutation manifest the clinical traits of the disease. Carriers are completely healthy but show a reduced mean cell volume (MCV) of red blood cells (Rosatelli et al. 1992), which is the variable usually used to identify them.

Beta-thalassemia is prevalent in Mediterranean countries, the Middle East, Central Asia, India, Southern China, and the Far East as well as countries along the north coast of Africa and in South America. The highest carrier frequency is reported in Cyprus, Sardinia, and Southeast Asia (Galanello and Origa 2010). In Sardinia, beta-thalassemia carriers make up about 15% of the population and a single mutation accounts for 95% of the beta-thalassemia mutations (Rosatelli et al. 1992; Cao et al. 2008).

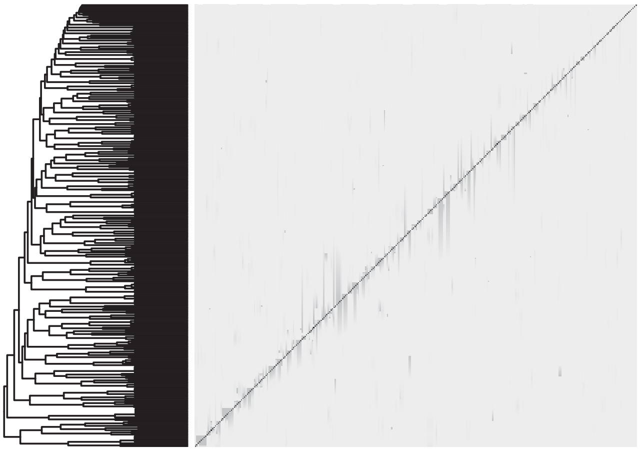

Figure 2. Dendrogram representing family tree (on the left) and kinship matrix (on the right) of all individuals in the beta-thalassemia dataset. Shading in the kinship matrix indicates the degree of relatedness between two individuals, where the darker the color the stronger the relationship between individuals. The kinship matrix defines a neighborhood structure in the population studied that can be naturally incorporated into the model.

The dataset studied here comes from Talana, a town in the province of Ogliastra, Sardinia, Italy. It is an isolated population characterized by a great deal of homogeneity in lifestyle and eating habits as well as a high endogamy and consanguinity. We had data on MCV, in logarithmic scale, of 306 related individuals originating from two common ancestors and from 6097 SNPs (more details on the dataset can be found in Cabras et al. 2011). The kinship matrix and the family tree in the sample are depicted in Figure 2.

We applied our two-stage procedure for GWAS for these data. We selected 129 SNPs in the first stage thus reducing the original dimension of the problem from 6097 to 129 SNPs. To assess the influence of the random effects in the SNP selection, we also estimated the regression model without the kinship data on the individuals in the sample. In this case, the number of selected SNPs was far greater, 271, showing that the inclusion of that kinship information into the model is relevant with regard to the efficiency of the model selection procedure. In the second stage, we used the selected 129 SNPs in the first stage and ran the random search algorithm 100 times, using 10,000 iterations for the Gibbs sampling in each run and starting from a different random initial model.

We also used a Bayesian lasso (BL) approach for model in Equation (4) and GEMMA software approach (Zhou, Carbonetto, and Stephens 2013) for comparison purposes. The parameters for the BL approach were estimated using the Gibbs sampling proposed in Park and Casella (2008) based on 5000 iterations. The penalty parameter was estimated using an empirical Bayes marginal maximum likelihood implemented in each Gibbs step using 500 simulations for each fixed parameter vector. Finally, a given SNP was selected by BL if the subsequent 95% credible interval did not contain the zero. GEMMA is the software implementing the Genome-wide Efficient Mixed Model Association algorithm (Zhou and Stephens 2012) and specifically, we used the Bayesian sparse linear mixed model that implements a spike and slab prior on regression coefficients

Table 1. Most frequently selected SNPs in the beta-thalassemia dataset, first, third quartile and median of the posterior distribution of the probability of inclusion; the last column is 1 or 0 whether the SNP is selected or not with BL.

SNP	N. times selected	$P^{\text{incl}}_{0.25}$	$P^{\text{incl}}_{0.5}$	$P^{\text{incl}}_{0.75}$	BL
rs10837540	99	1.00	1.00	1.00	1
rs11036238	98	1.00	1.00	1.00	1
rs4945957	97	1.00	1.00	1.00	1
rs4851857	97	0.99	1.00	1.00	1
rs10017809	94	0.98	1.00	1.00	1
rs9405550	94	0.96	0.99	1.00	1
rs13399239	90	0.96	0.99	1.00	1
rs544395	94	0.95	0.99	1.00	1
rs6822640	87	0.94	0.99	1.00	1
rs2173576	91	0.94	0.98	1.00	1
rs2055902	89	0.94	0.98	1.00	1
rs4553529	87	0.93	0.99	1.00	1
rs11047765	86	0.93	0.98	1.00	1
rs3112495	86	0.92	0.98	1.00	1
rs357116	87	0.92	0.98	1.00	1
rs9641113	86	0.91	0.98	1.00	1
rs3129877	88	0.91	0.99	1.00	1
rs1552484	90	0.90	0.97	1.00	0
rs12613635	87	0.88	0.97	1.00	0
rs6945778	84	0.87	0.98	1.00	1
rs17113771	84	0.87	0.98	1.00	1
rs181623	82	0.84	0.99	1.00	1
rs194528	81	0.82	0.98	1.00	1
rs6994583	84	0.82	0.96	1.00	0
rs6789065	83	0.81	0.96	1.00	1
rs9356011	83	0.79	0.97	1.00	1
rs2541389	82	0.77	0.98	1.00	1
rs4411225	79	0.76	0.95	1.00	1
rs699539	83	0.75	0.94	0.99	1
rs470014	81	0.73	0.95	1.00	1
rs17321742	83	0.69	0.96	1.00	1
rs1430620	78	0.68	0.95	1.00	1
rs17770069	79	0.68	0.98	1.00	1
rs4740004	78	0.65	0.97	1.00	1
rs6830552	77	0.64	0.95	0.99	1
rs6989470	76	0.61	0.93	0.99	0
rs8069352	79	0.58	0.92	1.00	0
rs2518110	76	0.56	0.96	1.00	0
rs6056536	76	0.54	0.91	0.98	0
rs7792551	75	0.52	0.91	0.99	1
rs1483460	75	0.52	0.90	0.99	0
rs7091802	75	0.52	0.97	1.00	1
rs1011969	75	0.51	0.94	0.99	0

NOTE: The highlighted SNPs are known to be related to the disease.

(George and McCulloch 1993; Ishwaran and Rao 2005) which also incorporates the kinship matrix for individual random effects (Zhou, Carbonetto, and Stephens 2013). The default setup and priors specified in Zhou, Carbonetto, and Stephens (2013) were used.

Code was implemented in R (R Core Team 2017) and the following libraries were also required: `BayesVarSel` (Garcia-Donato and Forte 2017), `kinship2` (Therneau and Sinnwell 2015), `INLA` (Rue, Martino, and Chopin 2009), `LearnBayes` (Albert 2014), `MCMCpack` (Martin, Quinn, and Park 2011), `plyr` (Wickham 2011), and `statmod` (Giner and Smyth 2016).

Table 1 summarizes the results of this second stage through the number of times (from the total of the 100 simulations) that each SNP is selected, the first, median and third quartile of the posterior distribution of the subsequent probability of inclusion described above, and the results of the BL selection. The table shows only those SNPs for which that first quartile is greater than 0.5. Remarkably, the two SNPs with highest probabilities are already known to be related to similar diseases: rs10837540 is mentioned in a specific GWAS beta-thalassemia study (Uda et al. 2008), while rs11036238 is located near the HBB gene which is directly related with hemoglobin and beta-thalassemia, and it has also been found to be related to malaria (Jallow et al. 2009). Results of the BL selection in the last column indicates whether the subsequent SNP was selected (value 1) or not (value 0). They indicate that 36 SNPs are related to the MCV variable, 34 of them are in Table 1. There is a great concordance between our results and those obtained from BL selection, mainly in the first 35 SNPs with higher values in the first quartile of the inclusion probabilities where there were only three discrepancies. The first 43 SNPs reported by GEMMA do not match with those reported by our approach or by BL, except for one SNP, rs8069352, which does not appear to be related to thalassemia. This could be due to the small sample available here (306 individuals) with respect to the 6097 SNPs analyzed, which again calls for a two-step procedure as proposed here.

3.2. *A Toy Example*

We discuss a toy example to exemplify the modeling features behind the two-stage method proposed and how it performs with respect to GEMMA and BL, focusing on the explanation of relevant scientific questions in easy terms.

Consider a simple genealogy tree with $n = 10$ individuals from two different families, as in Figure 3. A quantitative trait, y, for each individual was observed, as well as the number of dominant alleles in $p = 3$ SNPs. Figure 3 represents the structure of both families, parents, and children (three in both cases). The numerical information associated with each individual includes an identification number, his/her SNP information (a vector of dimension $p = 3$), and the value of the response variable, y. The gender of each individual is represented by a rectangle (male) or a circle (female).

Note that values of y for the family on the left are higher than the ones for the family on the right. The first SNP is strongly associated with the trait: 0 for all relatives in the first family and 2 for all members in the second. A similar situation occurs for the second SNP, always 0 in the second family, and 1 or 2 in the first one. SNP3 does not seem to be clearly related with the trait. In this example, we expect that the probability of association with the trait is higher in the case of SNPs 1 and 2 and lower for SNP3.

Table 2 shows the posterior inclusion probability for each SNP in each stage of our proposal as well as for the GEMMA software with the kinship information and all SNPs, and for the Bayesian lasso regression model with the two SNPs selected in

Table 2. Posterior probability of inclusion for SNP1, SNP2, and SNP3 from the first and second stage of our proposal, GEMMA software, and Bayesian lasso (BL) regression.

Model	SNP1	SNP2	SNP3
First stage	0.92	1.00	0.61
GEMMA	0.83	0.81	0.47
Second stage	0.78	0.49	
BL	0.96	0.91	

NOTE: Although all methods are able to select the related SNPs (1 and 2), our proposal provides a discrimination of the degree of that relationship (SNP1 more clearly related than SNP2).

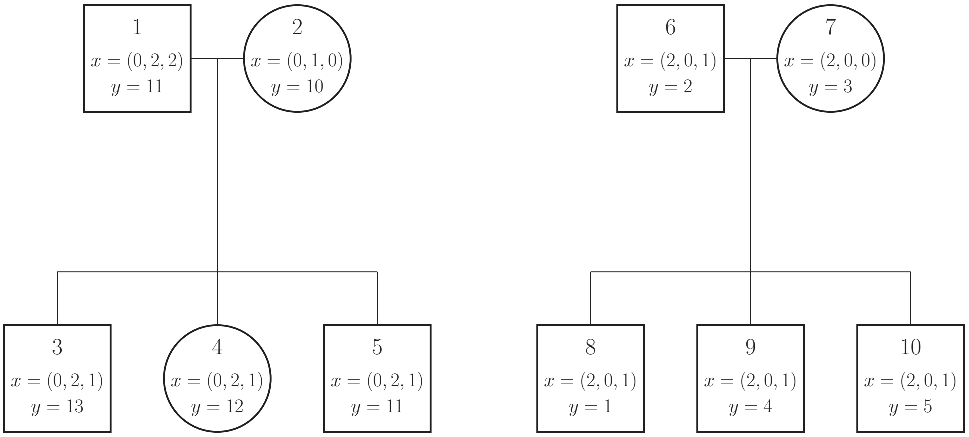

Figure 3. Family trees of the individuals in the toy example. The gender is represented by rectangles (male) and circles (female). The value of the three SNPs, x, and of the response variable, y, for each individual are indicated below each identification number. The family on the left is clearly affected by the disease whereas the family on the right is not. Only SNP1 and SNP2 are related to the disease.

our first stage. Our proposal, in the first stage, considered three independent regression models for explaining the trait. Each of the models includes the family tree information and the corresponding SNP information. Posterior 95% credible intervals discarded SNP3 for the second stage. Consequently, only SNP1 and SNP2 were included in the objective Bayes model selection process of the second stage, which considers jointly SNP1 and SNP2. It should be pointed out that, in the first stage, the posterior inclusion probability for SNP2 (1.00) is higher than the one corresponding to SNP1 as a result of the higher relevance of the kinship information in the presence of SNP1 than SNP2. Results from the second stage provided posterior inclusion probabilities 0.78 and 0.49 for SNP1 and SNP2, respectively, as expected. These are very conservative probabilities due to the small sample size which finally avoid large false discovery rates.

Results from GEMMA are similar to the ones obtained with the first stage in the sense that focus is on SNP1 and SPN2—the main relationships with the trait—but do not discriminate between them. Outcomes are not conclusive with regard to SNP3, showing a result close to 0.5. BL only applies to SNPs in the second stage and yields results in the same direction that ours with higher inclusion probabilities but without a clear distinction between them.

3.3. Comparing Approaches With a Simulation Study

We used some of the information from the beta-thalassemia dataset to conduct a simulation study to compare the Bayesian variable selection procedure (BVS) described in the second stage of our proposal with the Bayesian lasso (BL), with and without kinship information, and GEMMA software approaches. The Bayesian model implemented in GEMMA software always includes a kinship matrix (either provided or estimated) and thus a family effect. Therefore, the improvement in including the family effect is only assessed for BVS and BL. The objective of this design was twofold: to assess the effect of the family relationship in the first stage and the performance of our approach in the second stage with regard to Bayesian lasso regression and the GEMMA software.

Simulated data consisted of 100 replications of a regression dataset with a design matrix, a vector of regression coefficients, and a vector of 306 values of the response variable. In each replica, r, the design matrix $\boldsymbol{X}^{(r)}$ was defined from $p = 1000$ SNPs randomly selected from the original set of 6097 SNPs, the vector of regression coefficients $\boldsymbol{\beta}^{(r)}$ are all zero except for 10 SNPs randomly selected from the 1000 above whose coefficient values were randomly assigned from the set $\{-5, -2, 2, 5\}$, and the vector of response variable values generated from a normal distribution with vector of means $\boldsymbol{X}^{(r)}\boldsymbol{\beta}^{(r)}$ and variance-covariance matrix equal to the identity matrix.

The results of those analyses are presented in Table 3 through the sample mean and standard error of the empirical false discovery rate (FDR) and the false nonrejection rate (FNR) in 100 replications.

The modeling procedure which includes the family relationship among individuals in the sample has a shrinkage effect, improving the performance of the first stage. There are no differences among the methods with respect to the FNR, surely due to the large dimension of p. This is not the case for the FDR between BL and BVS: with and without familiar effects, differences in mean are lower for the BVS procedure. GEMMA

Table 3. Mean and standard error (in parenthesis) of the false discovery rate (FDR) and of the false nonrejection rate (FNR) produced by our proposal which includes the results of the two-stage procedure (first stage and second stage BVS in the table) with and without family information in the first stage, GEMMA software with family information, and Bayesian lasso (BL) regression which only uses the selected SNPs from the first stage, with and without family information.

Model	FDR	FNR
Family effect		
First stage	0.016 (8×10^{-4})	0.002 (1×10^{-4})
GEMMA	0.001 (1×10^{-3})	0.024 (5×10^{-3})
Second stage BVS	0.001 (2×10^{-4})	0.002 (1×10^{-4})
BL	0.003 (2×10^{-4})	0.002 (1×10^{-4})
No family effect		
First stage	0.058 (20×10^{-4})	0.003 (1×10^{-4})
Second stage BVS	0.007 (2×10^{-4})	0.003 (1×10^{-4})
BL	0.009 (5×10^{-4})	0.003 (1×10^{-4})

provides similar values of FDR and FNR, albeit with a larger variability, which may justify the differences obtained in the beta-thalassemia dataset.

4. Conclusions

We propose a two-stage approach for GWAS in which the family relationships between individuals are known. In the first stage, this information is included as a random effect in the regression model defining the relation between the response and each SNP. The promising SNPs selected in this stage are only considered in the second stage, which compares all possible models with the null model via BF to select the best model. As the space of all possible models is too large, a random search strategy is used for estimating the inclusion probabilities for each SNP.

The inclusion of the family relationship in the data by a random effect modeled with a Gaussian Markov random field has a shrinkage effect, as it is shown in the results of a simulation study. The lower FDR indicates that it facilitates a greater dimension reduction and a finer SNP selection. Additionally, in light of the results shown, our approach seems to be more effective in model selection than the Bayesian lasso.

We only use kinship information in the first stage but it could also be incorporated into the second one (as in Yazdani and Dunson 2015). It surely depends on the particular study analyzed: our benchmark study dealt with human populations and despite the fact that the family information was relevant it was not very strong thus producing identifiability problems. This is not the case in the article by Yazdani and Dunson (2015) within the framework of animal breeding, with a strong pedigree structure. Additionally, although the relatively moderate number of SNPs is shown in the example, the model proposed is valid for higher dimension problems.

A line of future work, for this kind of data, would be to use the familiar effect coupled with sparse priors as spike and slab (Ročková and George 2015) and/or nonlocal priors (Shin, Bhattacharya, and Johnson 2018) to approach the $p \gg n$ problem in a one-step analysis.

Supplementary Material

Appendix: Details on the implementation to enable readers to apply the article's proposed method to their own data. (supplem.pdf)

Acknowledgments

The authors thank the editor, associate editor, and referees for their useful comments that improved very much the original version of the article, and also Mario Pirastu and Maria Pina Concas for providing the data.

Funding

This article was partially funded by MTM2016-77501-P research grant and ECO2012-38442, RYC-2012-11455 projects from the Spanish Ministry of Economy and Competitiveness, and CRP-59903 from Regione Autonoma della Sardegna (Italy).

ORCID

Carmen Armero http://orcid.org/0000-0001-9839-6442
Stefano Cabras http://orcid.org/0000-0001-6690-8378
María Eugenia Castellanos http://orcid.org/0000-0001-7920-2307
Alicia Quirós http://orcid.org/0000-0001-5259-4793

References

Abecasis, G., Cherny, S., Cookson, W., and Cardon, L. (2002), "Merlin-Rapid Analysis of Dense Genetic Maps Using Sparse Gene Flow Trees," *Nature Genetics*, 30, 97–101. [1]

Albert, J. (2014), *LearnBayes: Functions for Learning Bayesian Inference, R Package Version 2.15*, available at *https://CRAN.R-project.org/package=LearnBayes*. [5]

Balding, D. J. (2006), "A Tutorial on Statistical Methods for Population Association Studies," *Nature*, 7, 781–791. [1]

Barbieri, M. M., and Berger, J. O. (2004), "Optimal Predictive Model Selection," *The Annals of Statistics*, 32, 870–897. [4]

Bayarri, M., Berger, J., Forte, A., and García-Donato, G. (2012), "Criteria for Bayesian Model Choice With Application to Variable Selection," *The Annals of Statistics*, 40, 1550–1577. [2,3]

Benyamin, B., Visscher, P. M., and McRae, A. F. (2009), "Family-Based Genome-Wide Association Studies," *Pharmacogenomics*, 10, 181–190. [1]

Berger, J. O., and Pericchi, L. R. (2001), "Objective Bayesian Methods for Model Selection: Introduction and Comparison," in *Model Selection* (Vol. 38), ed. P. Lahiri, Beachwood, OH: Institute of Mathematical Statistics, pp. 135–207 [2,3]

Cabras, S., Castellanos, M. E., Biino, G., Persico, I., Sassu, A., Casula, L., del Giacco, S., Bertolino, F., Pirastu, M., and Pirastu, N. (2011), "A Strategy Analysis for Genetic Association Studies With Known Inbreeding," *BMC Genetics*, 12, 63–74. [4]

Cao, A., Congiu, R., Sollaino, M., Desogus, M., Demartis, F., Loi, D., Cau, M., and Galanello, R. (2008), "Thalassemia and Glucose-6-Phosphate Dehydrogenase Screening in 13- to 14-Year-Old Students of the Sardinian Population: Preliminary Findings," *Community Genetics*, 11, 121–128. [4]

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015), "Bayesian Linear Regression With Sparse Priors," *The Annals of Statistics*, 43, 1986–2018. [1]

Eilers, P. H., and Marx, B. D. (1996), "Flexible Smoothing With b-Splines and Penalties," *Statistical Science*, 11, 89–102. [1]

Frank, L. E., and Friedman, J. H. (1993), "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, 35, 109–135. [1]

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33, 1–22. [1]

Fu, W. J. (1998), "Penalized Regressions: The Bridge Versus the Lasso," *Journal of Computational and Graphical Statistics*, 7, 397–416. [1]

Galanello, R., and Origa, R. (2010), "Beta-Thalassemia," *Orphanet Journal of Rare Diseases*, 5, 1–11. [4]

Garcia-Donato, G., and Forte, A. (2017), *BayesVarSel: Bayes Factors, Model Choice and Variable Selection in Linear Models, R Package Version 1.7.1*, available at *https://CRAN.R-project.org/package=BayesVarSel* [4,5]

García-Donato, G., and Martínez-Beneito, M. (2013), "On Sampling Strategies in Bayesian Variable Selection Problems With Large Model Spaces," *Journal of the American Statistical Association*, 108, 340–352. [2,4]

George, E. I., and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889. [1,5]

——— (1997), "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373. [2,4]

Giner, G., and Smyth, G. K. (2016), "statmod: Probability Calculations for the Inverse Gaussian Distribution," *R Journal*, 8, 339–351. [5]

Girosi, F., Jones, M., and Poggio, T. (1993), "Priors Stabilizers and Basis Functions: From Regularization to Radial, Tensor and Additive Splines," C. B. C. L. Paper No. 75, Artificial Intelligence Laboratory Massachusetts Institute of Technology. [1]

Gradshteyn, I. S., and Ryzhi, I. M. (1965), *Table of Integrals, Series and Products*, Boston, MA: Academic Press Inc. [3]

Herold, C., Hooli, B. V., Mullin, K., Liu, T., Roehr, J. T., Mattheisen, M., Parrado, A., Bertram, L., Lange, C., and Tanzi, R. E. (2016), "Family-Based Association Analyses of Imputed Genotypes Reveal Genome-Wide Significant Association of Alzheimer's Disease With Osbpl6, ptprg and pdcl3," *Molecular Psychiatry*, 21, 1608–1612. [1]

Hoerl, A. E., and Kennard, R. W. (1988), "Ridge Regression," in *Encyclopedia of Statistical Sciences* (Vol. 8), eds. N. L. Johnson, S. Kotz, and C. B. Read, New York: Wiley, pp. 129–136. [1]

Ishwaran, H., and Rao, J. S. (2005), "Spike and Slab Variable Selection: Frequentist and Bayesian Strategies," *Annals of Statistics*, 33, 730–773. [1,5]

Jallow, M., Teo, Y. Y., Small, K. S., Rockett, K. A., Deloukas, P., Clark, T. G., Kivinen, K., Bojang, K. A., Conway, D. J., Pinder, M., Sirugo, G., Sisay-Joof, F., Usen, S., Auburn, S., Bumpstead, S. J., Campino, S., Coffey, A., Dunham, A., Fry, A. E., Green, A., Gwilliam, R., Hunt, S. E., Inouye, M., Jeffreys, A. E., Mendy, A., Palotie, A., Potter, S., Ragoussis, J., Rogers, J., Rowlands, K., Somaskantharajah, E., Whittaker, P., Widden, C., Donnelly, P., Howie, B., Marchini, J., Morris, A., SanJoaquin, M., Achidi, E. A., Agbenyega, T., Allen, A., Amodu, O., Corran, P., Djimde, A., Dolo, A., Doumbo, O. K., Drakeley, C., Dunstan, S., Evans, J., Farrar, J., Fernando, D., Hien, T. T., Horstmann, R. D., Ibrahim, M., Karunaweera, N., Kokwaro, G., Koram, K. A., Lemnge, M., Makani, J., Marsh, K., Michon, P., Modiano, D., Molyneux, M. E., Mueller, I., Parker, M., Peshu, N., Plowe, C. V., Puijalon, O., Reeder, J., Reyburn, H., Riley, E. M., Sakuntabhai, A., Singhasivanon, P., Sirima, S., Tall, A., Taylor, T. E., Thera, M., Troye-Blomberg, M., Williams, T. N., Wilson, M., Kwiatkowski, D. P., Wellcome Trust Case Control Consortium, and Malaria Genomic Epidemiology Network. (2009), "Genome-Wide and Fine-Resolution Association Analysis of Malaria in West Africa," *Nature Genetics*, 41, 657–665. [5]

Kass, R., and Raftery, A. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795. [2]

Lee, A., Caron, F., Doucet, A., and Holmes, C. (2012), "Bayesian Sparsity-Path-Analysis of Genetic Association Signal Using Generalized t Priors," *Statistical Applications in Genetics and Molecular Biology*, 11, Article 5. [2]

Li, Q., and Lin, N. (2010), "The Bayesian Elastic Net," *Bayesian Analysis*, 5, 151–170. [1]

Malecot, G. (1948), *Les Mathematiques de l'Heredire*, Paris: Masson et Cie. [2]

Martin, A. D., Quinn, K. M., and Park, J. H. (2011), "MCMCpack: Markov Chain Monte Carlo in R," *Journal of Statistical Software*, 42, 22. Available at *http://www.jstatsoft.org/v42/i09/* [5]

Nychka, D. W. (2000), "Spatial-Process Estimates as Smoothers," in *Smoothing and Regression: Approaches, Computation, and Application*, ed. M. G. Schimek, New York: Wiley, pp. 393–424. [1]

O'Sullivan, F. (1986), "A Statistical Perspective on Ill-Posed Inverse Problems," *Statistical Science*, 1, 502–518. [1]

Ott, J., Kamatani, Y., and Lathrop, M. (2011), "Family-Based Designs for Genome-Wide Association Studies," *Nature Reviews Genetics*, 12, 465–474. [1]

Park, T., and Casella, G. (2008), "The Bayesian Lasso," *Journal of the American Statistical Association*, 103, 681–686. [1,4]

R Core Team (2017), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. [5]

Ročková, V., and George, E. I. (2015), "The Spike-and-Slab Lasso," *Journal of the American Statistical Association*, 113, 431–444. [7]

Rosatelli, C., Leoni, G., Tuveri, T., Scalas, M. T., Mosca, A., Galanello, R., Gasperini, D., and Cao, A. (1992), "Heterozygous Beta-Thalassemia: Relationship Between the Hematological Phenotype and the Type of Beta-Thalassemia Mutation," *American Journal of Hematology*, 39, 1–4. [4]

Rosatelli, M., Dozy, A., Faa, V., Meloni, A., Sardu, R., Saba, L., Kan, Y., and Cao, A. (1992), "Molecular Characterization of Beta-Thalassemia in the Sardinian Population," *American Journal of Human Genetics*, 50, 422–426. [4]

Rue, H., Martino, S., and Chopin, N. (2009), "Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations (with discussion)," *Journal of the Royal Statistical Society*, Series B, 71, 319–392. [3,5]

Ruppert, D., Wand, M., and Carroll, R. (2003), *Semiparametric Regression*(*Cambridge Series in Statistical and Probabilistic Mathematics*), Cambridge: Cambridge University Press. [1]

Scott, J., and Berger, J. (2010), "Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-Selection Problem," *The Annals of Statistics*, 38, 2587–2619. [2,3]

Shin, M., Bhattacharya, A., and Johnson, V. E. (2018), "Scalable Bayesian Variable Selection Using Nonlocal Prior Densities in Ultrahigh-Dimensional Settings," *Statistica Sinica*, 28, 1053–1078. [4,7]

Therneau, T. M., and Sinnwell, J. (2015), *kinship2: Pedigree Functions, R Package Version 1.6.4*, available at *https://CRAN.R-project.org/package=kinship2* [5]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [1]

Trecartin, R., Liebhaber, S., Chang, J., Lee, K., Kan, Y., Fubetta, M., Angius, A., and Cao, A. (1981), "Beta Zero Thalassemia in Sardinia is Caused by a Nonsense Mutation," *Journal of Clinical Investigations*, 68, 1012–1017. [4]

Uda, M., Galanello, R., Sanna, S., Lettre, G., Sankaran, V. G., Chen, W., Usala, G., Busonero, F., Maschio, A., Albai, G., Piras, MG., Sestu, N., Lai, S., Dei, M., Mulas, A., Crisponi, L., Naitza, S., Asunis, I., Deiana, M., Nagaraja, R., Perseu, L., Satta, S., Cipollina, M. D., Sollainoh, C., Moi, P., Hirschhorn, J. N., Orkin, S. H., Abecasis, G. R., Schlessinger, D., and Cao, A. (2008), "Genome-Wide Association Study Shows Bcl11A Associated With Persistent Fetal Hemoglobin and Amelioration of the Phenotype of β-Thalassemia," *Proceedings of the National Academy of Sciences*, 105, 1620–1625. [5]

Wagner, M. (2013), "Rare-Variant Genome-Wide Association Studies: A New Frontier in Genetic Analysis of Complex Traits," *Pharmacogenomics*, 14, 413–424. [1]

Wahba, G. (1990), *Spline Models for Observational Data*, Philadelphia, PA: SIAM. [1]

Wickham, H. (2011), "The Split-Apply-Combine Strategy for Data Analysis," *Journal of Statistical Software*, 40, 1–29. Available at *http://had.co.nz/plyr* [5]

Yazdani, A., and Dunson, B. (2015), "A Hybrid Bayesian Approach for Genome-Wide Association Studies on Related Individuals," *Bioinformatics*, 31, 3890–3896. [2,7]

Zhou, X., Carbonetto, P., and Stephens, M. (2013), "Polygenic Modelling With Bayesian Sparse Linear Mixed Models," *PLoS Genetics*, 9, e1003264. [2,4]

Zhou, X., and Stephens, M. (2012), "Genome-Wide Efficient Mixed-Model Analysis for Association Studies," *Nature Genetics*, 44, 821–824. [4]

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society*, Series B, 67, 301–320. [1]