Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Weinwurm et al., J Biomet Biostat 2013, 4:5
DOI: 10.4172/2155-6180.1000180

Research Article Open Access

The Effect of Linkage Disequilibrium on Bayesian Genome-wide Association Methods

Stephan Weinwurm[1], Johann Sölkner[1] and Patrik Waldmann[2]*

[1]*Division of Livestock Sciences, University of Natural Resources and Life Sciences, Vienna, Austria*
[2]*Division of Statistics, Department of Computer and Information Science, Linköping University, SE-581 83 Linköping, Sweden*

Abstract

The goal of genome-wide association studies (GWAS) is to identify the best subset of single-nucleotide polymorphisms (SNPs) that strongly influence a certain trait. State of the art GWAS comprise several thousand or even millions of SNPs, scored on a substantially lower number of individuals. Hence, the number of variables greatly exceeds the number of observations, which also is known as the $p \gg n$ problem.

This problem has been tackled by using Bayesian variable selection methods, for example stochastic search variable selection (SSVS) and Bayesian penalized regression methods (Bayesian lasso; BLA and Bayesian ridge regression; BRR). Even though the above mentioned approaches are capable of dealing with situations where $p \gg n$, it is also known that these methods experience problems when the predictor variables are correlated. The potential problem that linkage disequilibrium (LD) between SNPs can introduce is often ignored.

The main contribution of this study is to assess the performance of SSVS, BLA, BRR and a recently introduced method denoted hybrid correlation based search (hCBS) with respect to their ability to identify quantitative trait loci, where SNPs are partially highly correlated. Furthermore, each method's capability to predict phenotypes based on the selected SNPs and their computational demands are studied. Comparison is based upon three simulated datasets where the simulated phenotypes are assumed to be normally distributed.

Results indicate that all methods perform reasonably well with respect to true positive detections but often detect too many false positives on all datasets. As the heritability decreases, the Bayesian penalized regression methods are no longer able to detect any predictors because of shrinkage. Overall, BLA slightly outperformed the other methods and provided superior results in terms of highest true positive/ false positive ratio, but SSVS achieved the best properties on the real LD data.

Keywords: High dimensional genomics; SNPs; Correlated predictors; Stochastic search variable selection; Bayesian lasso; Bayesian ridge regression

Introduction

In order to understand the functionality of the genome and its products, a great number of genome-wide association studies (GWAS) have been conducted [1,2]. The majority of GWAS tries to find statistical relations between single-nucleotide polymorphisms (SNPs), distributed over large parts of the genome, and variations of the phenotypes under consideration [3,4]. As of October 2013, 11,680 SNPs in 1,724 studies have been reported to be associated to certain phenotypes in humans [5]. The number of studies conducted on animals and plants is steadily increasing [6,7].

Typical datasets in GWAS are comprised of many thousands, sometimes millions, of SNPs sequenced from hundreds to tens of thousands of individuals. Because the number of variables greatly exceeds the number of observations in these data sets, it leads to a statistical difficulty often referred to as the $p \gg n$ problem [8]. One consequence of the $p \gg n$ problem is that standard multiple-regression becomes infeasible. Many methods have been suggested as solutions to this problem. Generally, these methods can be categorized into two main directions, repeated single-SNP regression with adjustment of significance thresholds or multiple SNP regression combined with dimension reduction or regularization.

Most of the GWAS carried out previously are single-SNP studies where each SNP is tested individually for its association to the phenotype [3,9]. The repeated single variable regression is easy to implement and computationally straightforward. However, it is necessary to adjust significance thresholds for multiple comparisons, and for that there is no general consensus on the best approach. It is well-known that family-wise error rate (FWER) procedures (e.g. Bonferroni correction) are too conservative. The false discovery rate (FDR) procedures seem to have better statistical properties in the $p \gg n$ setting, especially versions depending on empirical estimation of the null-distributions [10]. However, it has been argued that these approaches are too simple to elucidate the comprehensive architecture of the genome [11], and that SNPs with small individual effects, that would be regarded as non-significant in single SNP analyses, still can have a large influence collectively on a phenotype [12].

Recently, in order to identify more complex relationships, a shift to more sophisticated multi-SNP approaches has taken place [4,13]. Since most SNPs in GWAS can be assumed to have no influence on the phenotype, the methods are often formulated as variable or model selection for finding the optimal model from all possible models. In frequentist statistics penalized likelihood methods [14] have become

***Corresponding author:** Patrik Waldmann, Division of Statistics, Department of Computer and Information Science, Linköping University, SE-581 83 Linköping, Sweden, Tel: (+46)13-281000; Fax: (+46) 13 142231; E-mail: Patrik.Waldmann@liu.se

Received November 05, 2013; **Accepted** November 23, 2013; **Published** November 28, 2013

Citation: Weinwurm S, Sölkner J, Waldmann P (2013) The Effect of Linkage Disequilibrium on Bayesian Genome-wide Association Methods. J Biomet Biostat 4: 180. doi:10.4172/2155-6180.1000180

popular because of their capability of simultaneously selecting important variables and estimating their effects in high dimensional statistical inference [15].

Bayesian inference provides a sophisticated approach to high dimensional problems due to their ability to incorporate prior knowledge and their unified probabilistic approach of data analysis [16,17]. Furthermore, Bayesian regularization methods give easily interpretable and comparable results along with valid standard errors [18]. Unfortunately, Bayesian formulations of multi-SNP methods are generally demanding, necessitating increased computational power and therefore it is important to capitalize on their sparsity [19].

Bayesian methods for variable selection can be broadly divided into methods based on spike and slab mixture priors and regularization priors [20]. Stochastic search variable selection (SSVS) is a popular Bayesian variable selection method based on mixture modeling of two normal distributions for the regression coefficients and was introduced by George and McCulloch [21]. SSVS has been further developed in numerous studies and applied to genomic datasets in work by Yi et al. [22], Srivastava and Chen [23], Guan and Stephens [24], Chen et al. [25] and Skarman et al. [26]. Bayesian regularization is based on one continuous prior that resembles the spike and slab shape of the mixture priors. Both the Bayesian lasso (BLA) and Bayesian ridge regression (BRR) are based on exponential power priors, with BLA having a Laplace distribution and BRR a Gaussian distribution [18]. The Bayesian adoptions of the lasso and ridge regression offer the advantage of providing estimates for the parameters along with valid standard errors that can be used for variable selection [18]. Bayesian penalized regression was used in genomic studies for example by Yi and Xu [27], Li et al. [11] and Cai et al. [28]. The Bayesian lasso was used by Silva et al. [29] for genome-wide selection (i.e. prediction of genomic breeding values).

One often over-looked challenge for GWAS arises from indirect linkage disequilibrium (LD), i.e. causal SNPs being correlated with other SNPs [2]. It has been suggested that many previously unveiled SNPs are unlikely to be the real variations in the genome, simply due to the fact that proximate SNPs on the chromosome are often in LD and acting as so called proxies [30]. It has been shown that correlations between explanatory variables can influence Bayesian variable selection methods considerably [31]. In order to investigate these issues, both direct and indirect associations are considered in this work and more details are given in the Methods and Results section.

Liang and Keleman [32] investigated approaches for correlated datasets for complex diseases, whereas Li and Zhang [33] proposed Bayesian variable selection methods for high-dimensional genomic dataset with strong correlation between markers. Recently, Kwon et al. [34] developed an improved version of SSVS, named hybrid correlation-based search (hCBS), designed to incorporate correlations between predictors.

The work at hand assesses the variable selection and predictive performance of various Bayesian multi-SNP approaches, including SSVS [21], hCBS [34] as well as the BLA and BRR [18] applied to $p \gg n$ datasets with known patterns of LD. Gibbs sampling is used to obtain samples from the posterior distributions to identify the most promising models. Evaluation of the methods is based on three datasets. Two datasets contain a simulated correlation structure to investigate common patterns of LD often present in GWAS datasets and the third dataset is divided into four different scenarios and uses the LD pattern of a real genomic dataset. The study focuses on quantitative trait loci (QTL) where phenotypes can be measured on a continuous scale (i.e. normally distributed), for example, height or weight.

The paper is organized as follows: the Methods section outlines the methods under consideration including SSVS and hCBS as well as BLA and BRR and introduces the datasets used for evaluation. The Results section summarizes the results of the application of each method to two simulated datasets in the Block-wise correlation and Exponential decay correlation function (LD) subsection and four different kinds of datasets based on a real genomic SNP data in the Real chromosome data subsection. Finally, the paper closes with the Discussion section. The computational resources were provided by the Vienna Scientific Cluster.

Methods and Data

Bayesian methods

Multiple linear regression: Assume a linear and independent contribution of every causal SNP [3]. A multivariate linear regression model with n observations (individuals) and p predictors (SNPs) can be written as

$$Y = \mathbf{X}\beta + \varepsilon, \tag{1}$$

where the unit standardized predictor variables $x_{i,j}$ are collected into matrix $\mathbf{X}=\{X_1, X_2,\ldots, X_p\}$ of dimension $n \times p$ and the unit standardized continuous phenotypes y_i into vector Y of length n. The strength of the influence of each SNP on the phenotype is represented by its regression coefficients in vector $\beta=\{\beta_1, \beta_2,\ldots, \beta_p\}$. The last term, ε, denotes the error term which is assumed to be normally distributed with mean 0 and variance σ_ε^2. Note that (1) doesn't include a mean because of the standardization of $\mathbf{X}$ and Y.

Stochastic Search Variable Selection (SSVS): The original stochastic search variable selection (SSVS) was proposed by George and McCulloch [21] and further extended by George and McCulloch [35], Brown et al. [36] and Brown et al. [37]. SSVS is a Bayesian method that randomly explores a fraction of all possible models. This is done by introducing a latent indicator vector γ that indicates which predictors are included in and excluded from the current model by setting $\boldsymbol{\gamma}_j=\mathbf{1}$ and to $\boldsymbol{\gamma}_j=\mathbf{0}$, respectively. Equation (1) now becomes

$$Y=\mathbf{X}_\gamma\beta_\gamma+\varepsilon. \tag{2}$$

Since SSVS is a Bayesian method, the parameters of the regression model need to be assigned prior distributions.

$$\sigma_\varepsilon^2 \sim InvGamma\left(\frac{a}{2}, \frac{ab}{2}\right), \tag{3}$$

Where $a/2$ is the shape parameter and $ab/2$ is the scale parameter. Furthermore, a conjugate prior for the regression coefficients β_γ is used

$$\beta_\gamma \left| \gamma, \sigma_\varepsilon^2 \sim N\left(0_p, \sigma_\varepsilon^2 \mathbf{H}_\gamma\right), \right. \tag{4}$$

Where 0_p is a zero vector and $\mathbf{H}_\gamma$ can be seen as a penalty term for inclusion of variables. For $\mathbf{H}_\gamma$ the independent prior cI_{p_γ} is used, since it is computationally more favorable than the g-prior [38]. George and McCulloch [21] introduced a widely adopted prior for γ taking the form of an independent Bernoulli distribution

$$p(\gamma) = \omega^{p_\gamma}(1-\omega)^{p-p_\gamma} \tag{5}$$

where p_γ denotes the number of variables currently selected into the

subset; $p_\gamma = \sum_{j=1}^{p} \gamma_j$;

ω is considered as a prior assumption of the subset size, more specifically the ratio of variables included into the selected subset to the total number of variables [21]. In the majority of GWAS the number of relevant SNPs associated has been shown to be rather small [30]. Hence, ω can be set to $10/p$. Based on the specified priors, the marginal posterior model probability becomes

$$p(\gamma|\mathbf{X},Y) \propto g(\gamma) = \left|I_n + \mathbf{X}_\gamma \mathbf{H}_\gamma \mathbf{X}'_\gamma\right|^{-1/2} \left|\mathrm{Q}_\gamma\right|^{-(a+n)/2} p(\gamma), \tag{6}$$

where $\mathbf{Q}_\gamma = ab + Y'\left(I_n - \mathbf{X}_\gamma \mathbf{K}_\gamma^{-1} \mathbf{X}'_\gamma\right) Y$ and $\mathbf{K}_\gamma = \mathbf{X}'_\gamma \mathbf{X}_\gamma + \mathbf{H}_\gamma^{-1}$.

Since the posterior distribution has to be evaluated for 2^p different models in order to find the model with the highest posterior distribution, computation becomes infeasible with the large values of p that is typical for GWAS. Instead, Markov chain Monte Carlo methods are used to approximate the posterior distribution. Still, sampling through all predictors in every iteration of the Markov chain becomes prohibitive when the number of p becomes larger than a few thousands. Brown et al. [38] suggested an approach where a new vector γ^* is created from the current γ by either adding or removing a randomly chosen predictor with probability $q\omega$. With probability 1-$q\omega$, swap two predictors by choosing independently at random a 0 and a 1 in γ and altering both of them. This results in the following proposal distribution

$$q(\gamma^*|\gamma) = \begin{cases} \dfrac{\varphi}{p_\gamma}, \text{if} \left|p_\gamma - p_{\gamma^*}\right| = 1 \\ \dfrac{1-\varphi}{p_\gamma (p - p_\gamma)}, \text{if} \left|p_\gamma - p_{\gamma^*}\right| = 0. \end{cases} \tag{7}$$

Using the Metropolis algorithm, a new model is then accepted with probability

$$\min\left\{\frac{g(\gamma^*)}{g(\gamma)}, 1\right\}, \tag{8}$$

where $q\omega$ usually is set to 0.5. Hence, with these priors SSVS does not incorporate any information about the relationships between variables in the generation of a new candidate model.

Correlation-based Search (CBS): The correlation-based search (CBS) uses a similar approach as SSVS, except that CBS does not consider every variable as independent [34]. As previously outlined, genomic data often show high correlations due to LD present in the genome. Not considering correlation during variable selection can result in the inclusion of highly correlated variables at the cost of variables being not considered which are part of the true underlying subset. While SSVS chooses the variables for altering γ randomly, CBS considers the correlation between variables in every iteration of the Markov chain to propose an altered subset γ^*. Only variables having a low correlation are added to the current subset, whereas highly correlated variables are excluded from the current subset [34]. Compared to Brown et al. [38], components of γ are no longer independent Bernoulli variables and therefore the prior is modified to

$$p(\gamma) \propto \binom{p}{p_\gamma}^{-1} \frac{1}{p_\gamma} \tag{9}$$

Consequently, since the proposal of a new subset is no longer symmetrical, the proposal distribution is altered as well

$$q(\gamma^*|\gamma) = \begin{cases} \dfrac{\varphi}{2p_\gamma}, \text{if} \left|p_\gamma - p_{\gamma^*}\right| = 1 \\ \dfrac{1-\varphi}{p_\gamma}, \text{if} \left|p_\gamma - p_{\gamma^*}\right| = 0 \end{cases} \tag{10}$$

Because the proposal distribution is not symmetrical, CBS uses a Metropolis-Hastings algorithm to generate the Markov chain

$$\min\left\{\frac{g(\gamma^*) \big/ q(\gamma^*|\gamma)}{g(\gamma) \big/ q(\gamma|\gamma^*)}, 1\right\}. \tag{11}$$

Hybrid Correlation-based Search (HCBS): Hybrid correlation-based search (hCBS) is an iterative stochastic search method that randomly alternates between SSVS and CBS. Kwon et al. [34] suggested 90% CBS and 10% SSVS iterations for the construction of a Markov chain. In order to obtain only SSVS samples the CBS part is set to 0 to result in only SSVS moves in each iteration.

Bayesian Lasso (BLA): It was noted already by Tibshirani [14] that lasso estimates of the regression coefficients could be interpreted as the Bayes posterior mode under independent Laplace (double-exponential) priors. However, a Gibbs sampler for a full hierarchical model of the Bayesian lasso (BLA) was only recently introduced by Kyung et al. [18]. Consider the multiple regression model in (1) with standardized predictors and response. A hierarchical Bayesian model can be formulated as

$$\begin{aligned} & y|\mathbf{X},\beta,\sigma_\varepsilon^2 \sim N\left(\mathbf{X}\beta, \sigma_\varepsilon^2 I_n\right) \\ & \beta|\sigma_\varepsilon^2,\tau_1^2,\tau_2^2,\ldots\tau_p^2 \sim N\left(0_p, \sigma_\varepsilon^2 \mathbf{D}_\tau\right) \\ & \mathbf{D}_\tau = \mathrm{diag}\left(\tau_1^2,\tau_2^2,\ldots\tau_p^2\right) \\ & \tau_1^2,\tau_2^2,\ldots\tau_p^2 \sim \prod_{j=1}^{p} \frac{\lambda^2}{2} e^{-\lambda^2\tau_j^2/2} d\tau_j^2 \\ & \sigma_\varepsilon^2 \sim \frac{1}{\sigma_\varepsilon^2}, \end{aligned} \tag{12}$$

Where λ is the regularization parameter and $\tau_1^2,\tau_2^2,\ldots\tau_p^2$ are predictor specific variances. After integrating out $\tau_1^2,\tau_2^2,\ldots\tau_p^2$, the prior on β will have the form of a conditional Laplace (double exponential) prior

$$p\left(\beta|\sigma_\varepsilon^2\right) = \prod_{j=1}^{p} \frac{\lambda}{2\sigma_\varepsilon} e^{-\lambda|\beta_j|/\sigma_\varepsilon} \tag{13}$$

which ensures unimodality of β. Conveniently, λ is also assigned a hyperprior and thus it is not necessary to a priori estimate the appropriate amount of shrinkage by for example cross-validation or pilot runs of MCMC. Kyung et al. [18] show that a Gibbs sampler can be constructed by sampling from the following full conditional posterior distributions

$$\beta\left|\sigma_\varepsilon^2,\tau_1^2,\tau_2^2,\ldots\tau_p^2,\mathbf{X},y,\lambda\sim\ N\left(\left(\mathbf{X}'\mathbf{X}+\mathbf{D}_\tau^{-1}\right)^{-1}\mathbf{X}'y,\sigma_\varepsilon^2\left(\mathbf{X}'\mathbf{X}+\mathbf{D}_\tau^{-1}\right)^{-1}\right)\right.$$

$$\sigma_\varepsilon^2\left|\beta,\tau_1^2,\tau_2^2,\ldots\tau_p^2,\mathbf{X},y,\lambda\sim\ InvGamma\left(\frac{n-1+p}{2},\frac{1}{2}(y-\mathbf{X}\beta)'(y-\mathbf{X}\beta)+\frac{\lambda}{2}\beta'\mathbf{D}_\tau^{-1}\beta\right)\right.$$

$$1/\tau_j^2\left|\beta,\sigma_\varepsilon^2,\mathbf{X},y,\lambda\sim\ InvGaussian\left(\frac{\lambda^2\sigma_\varepsilon}{|\beta_j|},\lambda^2\right)I\left(\frac{1}{\tau_j^2}>0\right)\right. \quad (14)$$

$$\lambda^2\left|\beta,\sigma_\varepsilon^2,\tau_1^2,\tau_2^2,\ldots\tau_p^2,\mathbf{X},y\sim\ Gamma\left(p+r,\sum_{j=1}^p\tau_j^2/2+\delta\right)\right.$$

Where r=1 and δ=0.1.

Bayesian Ridge Regression (BRR): The Bayesian ridge regression (BRR) performs regularization by assigning the regression coefficients a Gaussian prior. The prior (13) is now modified to

$$p\left(\beta\left|\sigma_\varepsilon^2\right.\right)=\prod_{j=1}^p\frac{\lambda}{2\sigma_\varepsilon}e^{-\lambda\left|\beta_j\right|/\sigma_\varepsilon}. \quad (15)$$

However, the same hierarchical setup as for BLA is used to represent the BRR but with the modification τ_j^2 of the priors on $\tau_1^2,\tau_2^2,\ldots\tau_p^2$ and λ. According to Kyung et al. [38], the hierarchical lasso is adapted for ridge regression by giving all τ_j^2's a degenerative distribution at the same constant value

$$1/\tau_j^2=0+\lambda \quad (16)$$

and the regularization parameter is sampled from

$$\lambda\left|\beta,\sigma_\varepsilon^2,\tau_1^2,\tau_2^2,\ldots\tau_p^2,X,y\sim Gamma\left(\frac{p}{2}+r,\frac{1}{2\sigma_\varepsilon^2}\sum_{j=1}^p\beta_j^2+\delta\right)\right. . \quad (17)$$

Variable selection in bayesian regularization: Since neither the Bayesian lasso nor the Bayesian ridge regression is able to effectively set the regression coefficients of irrelevant variables exactly to zero, subsequent variable selection is performed by using the credible interval (CI) criterion [39]. A variable is excluded if the credible interval of the regression coefficient β_i covers zero. Consequently, a variable is considered relevant if zero lies outside of the credible interval. A common choice is a 95% credible interval. Li and Lin [39] argued that a 95% interval leads to too few selections. Hence, we decided to use also 90% and 50% CIs for comparative purposes.

Simulated data

Evaluation is based on three different datasets with varying patterns of LD. The first two datasets contain an artificially created LD structure in order to investigate its affect on the variable selection performance of the methods. The last dataset uses the correlation structure from a real GWAS dataset and is divided into four different scenarios where the set of influential SNPs, and thus their interrelationships is varied, resulting in different direct and indirect associations between SNPs and the phenotype.

Exponential decay correlation function (LD): The first simulated dataset consists of a correlation structure between the predictor variables that decreases exponentially with distance between SNPs, resembling the correlation often found over chromosomes [40]. 5000 predictors (SNPs) and 500 responses (phenotypes) were used in the GWAS datasets. First, the correlation matrix Υ_X was generated with the following correlation function $\rho=0.9^{|i-j|}$. The 5000 predictors were then generated from a multivariate normal distribution with mean vector μ=0 and covariance matrix Υ_X. The response variable was constructed as the sum of the first 10 SNPs plus a random draw from the univariate normal distribution with mean μ=0 and standard deviation σ_ε=1. Similar simulated datasets are used by Hastie and Zou [41] and Li and Lin [39].

Block-wise correlation (LD): The second simulated dataset yields a block-wise correlation structure and mimics a block of proximate correlated SNPs influencing the phenotype beside a second large block of correlated variables having no direct influence. A similar dataset has been used by Kwon et al. [34]. This dataset also contains p=5,000 SNPs and n=500 phenotypes. The following correlation matrix Υ_X was used

$$\mathbf{Y}_X=\begin{pmatrix}\mathbf{Y}_{11} & \mathbf{Y}_{12}\\ \mathbf{Y}_{21} & \mathbf{Y}_{22}\end{pmatrix} \quad (18)$$

Where Υ_{11} is a 10×10 matrix, corresponding to the correlation between the predictors associated with the phenotype. Υ_{12} and Υ_{21} denote the correlation matrices between the 10 predictors associated with the response and the remaining 4990 predictors outside Υ_{11}. Consequently Υ_{22} represents a 4990×4990 block. The off-diagonal elements were chosen as Υ_{11}=0.85 and Υ_{22}=0.55 whereas the elements of the diagonals were set to 1. All elements of Υ_{12} and Υ_{21} were selected to be 0.45. Furthermore β_i=0.5 for i=1,…10 and β_i=0 for i=11,…,5000 are the mean vector. The 5000 predictors were then generated from a multivariate normal distribution with mean vector β and covariance matrix Υ_X. The response was generated as the linear contribution of the 10 associated variables together with errors from the univariate normal distribution with mean μ=0 and standard deviation $\sigma_\varepsilon^2=1.25$ datasets were generated in both the exponential decay and block structure settings. All datasets were normalized and standardized $\sum_{i=1}^n x_{ij}=0$, $\sum_{i=1}^n y_i=0$ and $\sum_{i=1}^n x_{ij}^2=1$ for j=1,…, p.

Real chromosome correlation (LD): In the exponential decay and block structure settings the correlation matrices were artificially created and mimic only certain idealized properties of the LD structure over chromosomes. In order to assess the performance of the methods under consideration when applied to a dataset having the correlation structure of a real genomic dataset, two chromosomes were used from Austrian Fleckvieh cattle genotyped with the Illumina bovine 54K SNP chip. The dataset consists of 4697 SNPs from chromosome 1 and chromosome 2 sequenced from 2122 bulls. LD between all SNPs was estimated based on the standard Pearson correlation coefficient yielding a correlation matrix Υ_{rxr} of size 4697×4697. Four different scenarios were created to investigate differences between associations of non-correlated and correlated SNPs within and outside LD blocks.

- For the first scenario (*Associated SNPs inside LD-Blocks*), 10 LD-blocks with an average absolute correlation value of 0.65 were identified on the chromosomes. In every block the SNP X_i^* yielding the highest average correlation to the other SNPs in each block was selected to be associated with the phenotype. The predictors were generated as $X\sim N(0_r,\Upsilon_{rxr})$ and the response as $Y\sim\sum X_i^*+N(0,\sigma_\varepsilon^2)$. Moreover, in order to examine the influence of different amounts of information content (i.e. heritability), the datasets in each of the four scenarios were generated using four different values for the error variance $\sigma_\varepsilon^2=1,5,20,50$. All datasets were centered and normalized as for the two earlier data sets.

- In the second scenario (*Associated SNPs inside and correlated SNPs outside of LD-blocks*), in addition to the previous setup, for each of the 10 SNPs within LD-blocks assigned to be associated to the

phenotype, the highest correlated SNP lying outside of the LD-blocks was also assigned to be associated with the response, resulting in 20 relevant SNPs in total. The average absolute correlation between the SNPs in LD-blocks and SNPs lying outside was 0.31; whereas the mean correlation between the SNPs outside the LD-blocks was 0.05.

- The third setup (*Associated SNPs inside and non-correlated SNPs outside of LD-Blocks*) contains, besides the 10 SNPs inside the LD-blocks as in the first setup, 10 additional SNPs for which an association to phenotype was imputed. These additional SNPs, all outside of LD-blocks, are chosen by using the least correlated SNP for each of the 10 associated SNPs inside the LD-blocks. The mean absolute correlation between each SNP inside a block and the least correlated SNP outside was 0.003.

- For the fourth setup (*Associated SNPs outside of LD-Blocks*), only the 10 least correlated SNPs outside of LD-blocks, which were added in the preceding scenario, were used to construct the phenotype. The correlation between these SNPs was on average 0.04.

Computational analysis: A direct comparison of the computational efficiency of all methods is rather unfair, since the main feature of the SSVS-based methods is to perform variable selection during computation so that only a subset of the variables is used in every iteration, whereas BLA and BRR compute all variables in every iteration and perform variable selection subsequent to the computation. As a consequence, the performance of SSVS-based methods depends on the number of relevant variables and on the choices for the hyperparameters and will be considerably faster for most GWAS [24]. Moreover, since samples from the Metropolis -Hastings-algorithm are generated and then either accepted or rejected, identical samples are included in chain, leading to an increased autocorrelation. As a consequence, more samples are needed to give meaningful results. For those reasons, the comparison was based on the time required for one iteration, averaged over 500 iterations and 5 repeats per method. All simulation and analyses were performed with Matlab.

Results

hCBS and SSVS were run for 1,000,000 iterations for the datasets with the exponentially decaying correlation function and the block-wise correlation structure. Of these, 5,000 iterations were discarded as burn-in period. BLA and BRR were run for 15,000 iterations with 1,000 iterations removed as burn-in from the MCMC chain. The scale reduction factor R<1.1 and the effective sample size (*neff*) approaching 10,000 were used as criteria for convergence of the MCMC chains.

Exponential decay correlation function (LD)

hCBS and SSVS were run with ω=10/5000 (the hyperprior for the number of expected true variables), a=3 and b=1 and $\mathbf{H}_\gamma = cI_{p_\gamma}$ with c=0.05 (the penalty term for inclusion of variables). The proportion of CBS to SSVS moves was set to 0.9 [34] and to 0 to obtain the pure SSVS. 0.5 was used as probability for the swap or inclusion/exclusion move. 25 datasets were analyzed to obtain average results.

Figure 1 summarizes the average number of true and false positive detections over the 25 data sets. Selections are based on 95% CIs for BLA and BRR, and a threshold of 0.5 for the posterior inclusion probability (PIP) in the case of hCBS and SSVS. BRR performed best in terms of true positive detections (10), but also yielded the largest number of false positive detections (2.280). BLA detected 9.783 true positive variables on average and was the only method including no false positives. SSVS performed better than hCBS in terms of true positive detections (8.160 vs. 7.720), which is a rather unexpected result, since the correlation in the dataset is up to 0.9. Overall, BLA performed best in terms of highest fit criterion, i.e., $FIT = -(\#ExpPos - \#TruePos) - \#FalsePos = -(10 - 9.783) - 0 = -0.217$

Figures 2 and 3 show examples from one simulated dataset of the regression coefficients obtained by BLA and BRR as well as the 95%, the 90% and the 50% CIs for the first 20 variables. It is worth noting that BRR results in more dependency between the regression coefficients of the variables and CIs of similar size because of the joint prior in (16).

The prediction error (PE) was obtained by using the regression coefficients obtained from each of the 25 datasets β^i to predict the responses for the remaining 24 datasets $\mathrm{PE}^i = \frac{1}{24}\sum_k^{25\setminus i}\left(\mathrm{X}\frac{k}{\gamma}\hat{\beta}\frac{i}{\gamma} - Y^k\right)^2$

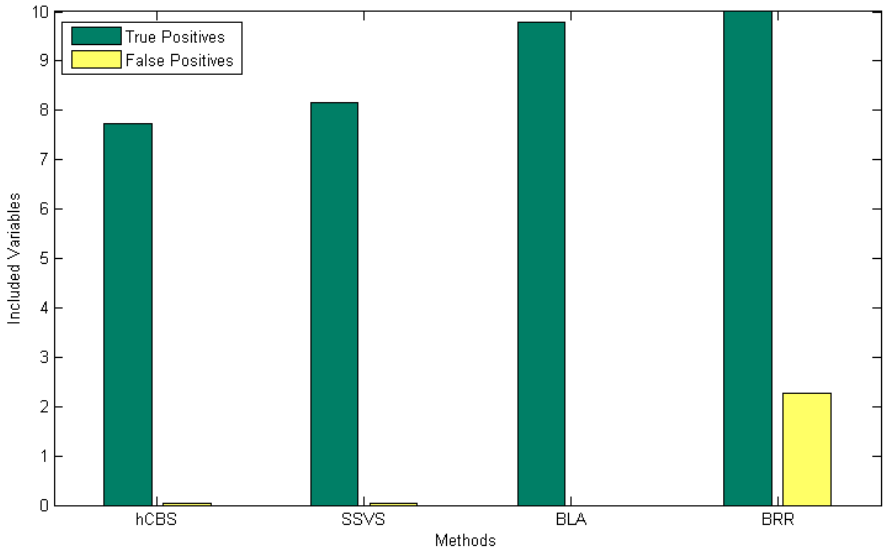

Figure 1: Average true and false positive detections by all methods over 25 exponentially decaying correlation function datasets. Selections are based on 95% CIs for BLA and BRR.

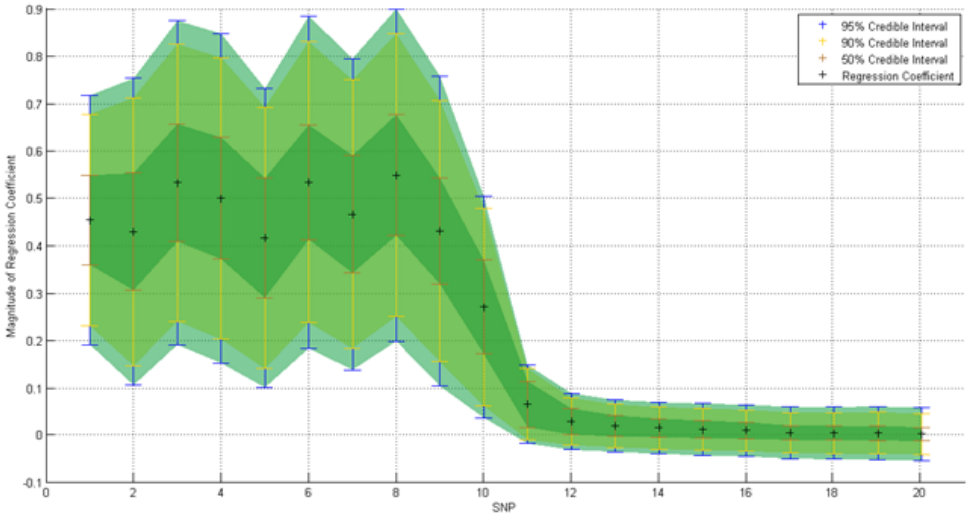

Figure 2: Mean regression coefficients and three different CIs of the 20 first predictors (SNPs) obtained by BLA when applied to one exponentially decaying correlation function data set.

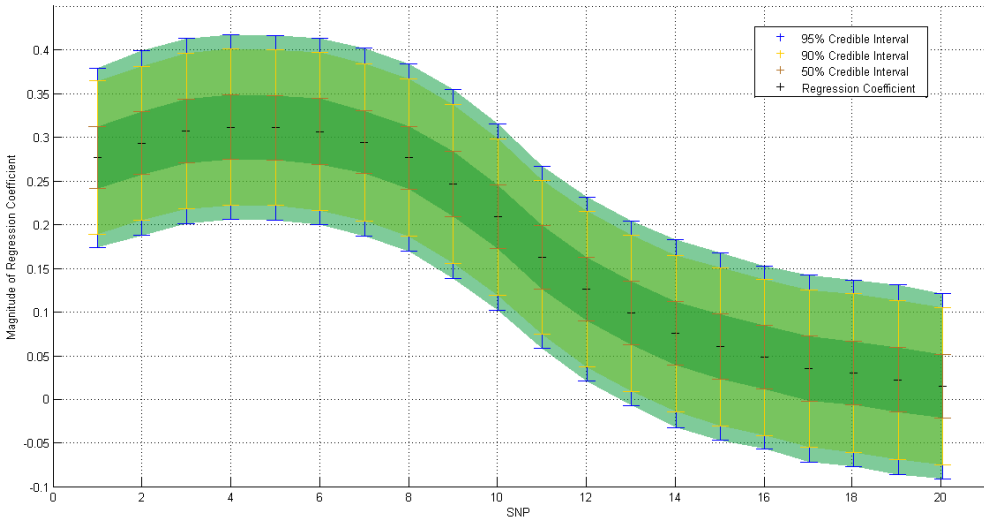

Figure 3: Mean regression coefficients and three different CIs of the 20 first predictors (SNPs) obtained by BRR when applied to one exponentially decaying correlation function data set.

and subsequently averaging over all PE^i computed. Table 1 shows the prediction errors obtained by the methods. As can be seen, all methods performed rather well with a similar prediction error.

The computational demands varied between the methods. hCBS and SSVS required less than 167 minutes for one data set, whereas BLA terminated after on average 20.8 hours and BRR took on average 22.1 hours.

Block-wise correlation (LD)

For hCBS and SSVS, the same hyperparameters were used as in the analyses of the exponentially decaying data, with the exception of c which was set to 1. Figure 4 shows the results using the same variable selection thresholds as for the exponentially decaying data. All methods, except for SSVS, were able to identify all relevant variables. BLA and BRR both performed very well and identified all associated SNPs with no false positive detections. hCBS found 10 true positive variables and 1.6 non-associated variables; whereas SSVS found on average 9.33 out of the 10 associated variables but yielded a lower number of false positives than hCBS.

Figure 5 shows an example based on one simulated dataset of the regression coefficients of the first 20 variables obtained by BLA along with the 95%, 90% and 50% CIs. Figure 6 depicts the corresponding plot for BRR. There it can be seen that for BRR the block-wise correlation structure results in less dependency between neighboring variables than the exponentially decaying correlation structure in Figure 3. The PE was estimated in the same way as for the exponentially decaying data sets. All methods performed similar in terms of PE with hCBS achieving slightly better results compared to the other methods. Results are shown in Table 1. Relaxing the variable selection criteria of PIP to 0.4 for SSVS and hCBS, and to 90% CIs for BLA and BRR resulted in a

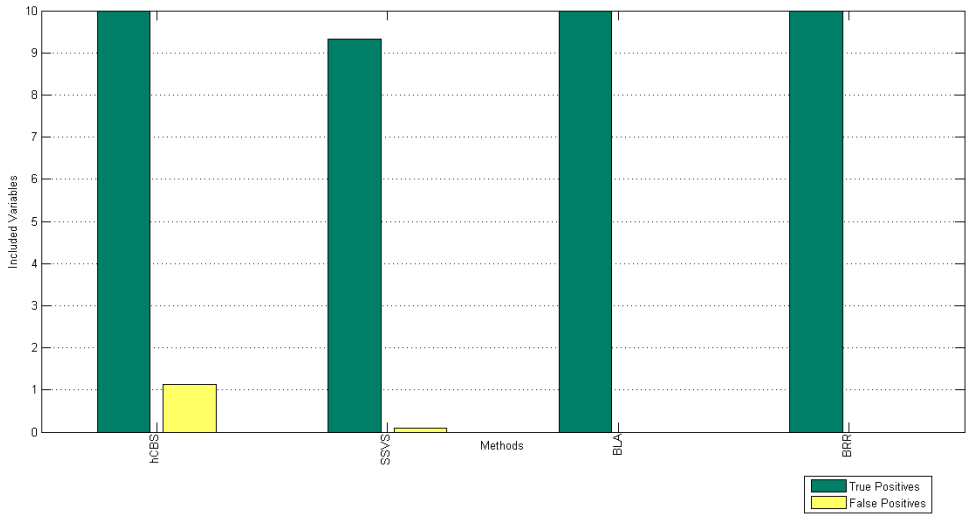

Figure 4: Average true and false positive detections by all methods over 25 block-wise correlation structure datasets. Selections are based on 95% CIs for BLA and BRR.

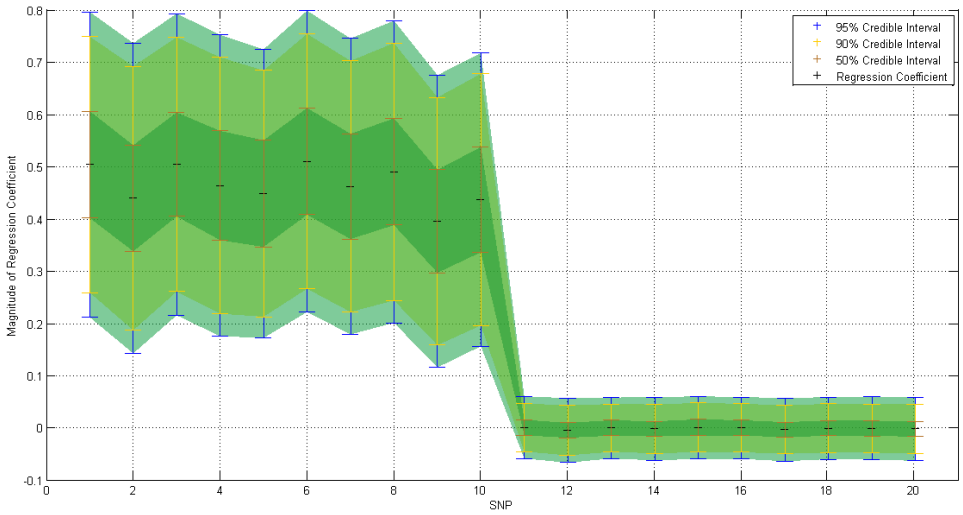

Figure 5: Mean regression coefficients and three different CIs of the 20 first predictors (SNPs) obtained by BLA when applied to one block-wise correlated data set.

Method	PE Exponentially decaying correlation function	PE Block-wise correlation structure
hCBS	1.19	1.07
SSVS	1.14	1.09
BLA	1.12	1.15
BRR	1.10	1.14

Table 1: Prediction error (PE) produced by each method for the exponentially decaying correlation function and the block-wise correlation data sets.

false positive rate of 3.7 for hCBS, whereas the number of false positive detections for the penalized regression methods remains unchanged. Changing the variable selection criterion for the Bayesian penalized regression methods to 50% yielded 427 and 677 false positives, leading to the conclusion that a 50% CIs is not restrictive enough to be used for variable selection purposes.

Computation of SSVS took on average 160 minutes and for hCBS slightly more than 167 minutes. 15,000 iterations of BLA and BRR required significantly more time. The former terminated after 17.75 hours and the latter finished after 22.14 hours, on average.

Real chromosome correlation (LD)

BRR and BLA were run for 40,000 iterations and both SSVS-based methods for 850,000 iterations. We first evaluated the effect of varying the hyperparameters of the SSVS-based methods on the *Associated SNPs inside LD-Blocks* data. For each of the $\sigma_\varepsilon^2 = \{1, 5, 20, 50\}$ settings, analyses were performed on all combinations between c={0.0001,0.001,0.01,0.1,1} and ω={0.005,0.01,0.02,0.04,0.1}. Hyperparameter a was set to 1 and b to its corresponding $\sigma_\varepsilon^2 = 1$ value. Convergence diagnosis (R and *neff*) and acceptance ratio were monitored as well as the number of true and false positives. The

	hCBS		SSVS		BLA				BRR			
σ_ε^2	TP	FP	TP	FP	TP95	FP95	TP90	FP90	TP95	P95	P90	FP90
1	8	2	10	0	7	1	10	10	10	38	10	56
5	8	2	9	0	1	0	4	0	3	4	7	16
20	7	3	7	0	0	0	1	0	0	0	1	0
50	5	2	4	0	0	0	0	0	0	0	0	0

Table 2: True and false positive detections by all methods when applied to the real cattle LD data (4697 SNPs from chromosome 1 and 2) with 10 SNPs inside LD-Blocks used to generate the phenotype. Four different values for the error variance σ_ε^2 were simulated.

	hCBS		SSVS		BLA				BRR			
σ_ε^2	TP	FP	TP	FP	TP95	FP95	TP90	FP90	TP95	P95	TP90	FP90
1	19	2	19	1	15	3	19	5	19	31	20	57
5	18	6	17	2	10	0	11	2	11	8	15	22
20	15	4	14	3	3	0	4	0	2	0	3	0
50	14	4	14	3	0	0	1	0	0	0	0	0

Table 3: True and false positive detections by all methods when applied to the real cattle LD data (4697 SNPs from chromosome 1 and 2) with 20 SNPs associated to the phenotype, 10 inside of LD-blocks and 10 correlated SNPs outside of LD-blocks. Four different values for the error variance σ_ε^2 were simulated.

	hCBS		SSVS		BLA				BRR			
σ_ε^2	TP	FP	TP	FP	TP95	FP95	TP90	FP90	TP95	P95	TP90	FP90
1	18	1	18	2	17	0	19	4	18	22	20	44
5	18	1	19	2	10	0	12	2	12	6	15	15
20	17	4	14	1	5	0	7	0	1	0	7	0
50	13	4	15	2	0	0	2	1	0	0	2	0

Table 4: True and false positive detections by all methods when applied to the real cattle LD data (4697 SNPs from chromosome 1 and 2) with 20 SNPs associated to the phenotype, 10 inside of LD-blocks and 10 non-correlated SNPs outside of LD-blocks. Four different values for the error variance σ_ε^2 were simulated.

complete list of results using various hyperparameters is available from the authors. The best average *FIT*-value over the four different error variances was obtained with ω=0.01 and c=1 for SSVS *FIT*=-2.5, and with ω=0.005 and c=1 for hCBS (*FIT*=-5.25). Based on these hyperarameters, SSVS identified all true positives only for the data set with highest information content $\sigma_\varepsilon^2 = 1$. hCBS was never able to identify all true positives. The best result was obtained for $\sigma_\varepsilon^2 = 1$ with 8 true positives. Moreover, hCBS reported more false positives than SSVS (Table 2). BRR performed better than BLA in terms of true positives but also identified considerably more false positives for datasets with high information content (Table 2). Both methods were unable to identify any true positives for $\sigma_\varepsilon^2 = 50$.

For the *Asscociated SNPs inside and correlated SNPs outside of LD-blocks*, the best average *FIT*-value over the four different error variances was achieved with ω=0.02 and c=1 for SSVS (*FIT*=-6.25), and with ω=0.01 and c=0.1 for hCBS (*FIT*=-7.5). hCBS and SSVS performed similar in terms of average true positive detections but SSVS detected considerably fewer false positives than hCBS (Table 3). BRR achieved the same number of true positives as hCBS and SSVS for $\sigma_\varepsilon^2 = 1$, but again included many more false positives. BLA was inferior regarding true positives by only including 15 of the 20 associated SNPs. When relaxing the variable selection to a 90% interval BLA includes 19 true positives and only 5 false positives. BRR found all associated SNPs but also selects 57 false positives. Similar to the previous example, both penalized regressions methods struggle with the detection of true positives in low information datasets such as $\sigma_\varepsilon^2 = 20, 50$.

The best average *FIT*-value over the four different error variances was obtained with ω=0.01 and c=0.1 for SSVS (*FIT*=-5.25), and with ω=0.005 and c=0.1 for hCBS (*FIT*=-6) for the *Associated SNPs inside and non-correlated SNPs outside of LD-Blocks* data. This time hCBS and SSVS performed similar with both including a high number of positive detections and relatively few false negatives. Table 4 summarizes the results obtained. The pattern of BLA and BRR not being able to detect a significant amount of true positives in low information datasets $(\sigma_\varepsilon^2 = 20, 50)$ is repeated. Furthermore, BLA again performs very well in terms of few false positives and BRR again selects too many false positives (Table 4).

As for the *Associated SNPs outside of LD-Blocks* data, the best average *FIT*-value over the four different error variances was obtained with ω=0.01 and c=1 for SSVS (*FIT*=0) and with ω=0.005 and c=0.1 for hCBS (*FIT*=-0.75). SSVS performed particularly well in this scenario as the method has been designed for datasets with little correlation between the predictors. SSVS manages to identify 10 out of 10 relevant variables with no false positives for all error variances (Table 5). hCBS also resulted in good true positive detection rates, but performed slightly inferior to SSVS in terms of false positives. BLA and BRR both have problems with the detection of true positives in low information datasets, especially for the data sets with $(\sigma_\varepsilon^2 = 50)$.

Computational analysis

From the results shown in Figure 7, it can be seen that the computation time of SSVS-based methods scale with increases in the number of phenotypes or more generally of observations. Note that the y-scale is logarithmic. In contrast, Bayesian penalized regression methods scale with the number of SNPs or in general with the number of variables. This stems from the fact, that in every iteration each variable has to be sampled separately which is the most time consuming operation. Bayesian penalized regression methods required on average 5.5 seconds per iteration for a dataset of size 5,000×5,000. In contrast SSVS-based methods took 1.7 seconds. If the data set consists of 500 phenotypes instead of 5,000, then Bayesian penalized regression methods still required around 5.5 seconds, whereas computation time of SSVS based methods decreased to 0.004 seconds. However, SSVS-based methods required many more iterations to be carried out than BLA and BRR.

Discussion

The main contribution of this study is to conduct a detailed comparison between the Bayesian penalized regression methods, Bayesian lasso (BLA) and Bayesian ridge regression (BRR), as well as stochastic search variable selection (SSVS) and hybrid correlation-based search (hCBS) on simulated phenotype data with real and simulated linkage disequilibrium (LD) between SNPs. The simulated datasets mimic certain properties of datasets common in genome-wide association studies (GWAS), such as high block-wise and exponentially

	hCBS		SSVS		BLA				BRR			
σ_ε^2	TP	FP	TP	FP	TP95	FP95	TP90	FP90	TP95	P95	TP90	FP90
1	10	0	10	0	10	0	10	0	10	9	10	19
5	10	1	10	0	10	0	10	1	10	2	10	5
20	10	1	10	0	7	0	7	0	6	0	7	1
50	10	1	10	0	0	0	0	0	0	0	0	0

Table 5: True and false positive detections by all methods when applied to the real cattle LD data (4697 SNPs from chromosome 1 and 2) with 10 SNPs outside of LD-blocks used to generate the phenotype. Four different values for the error variance σ_ε^2 were simulated.

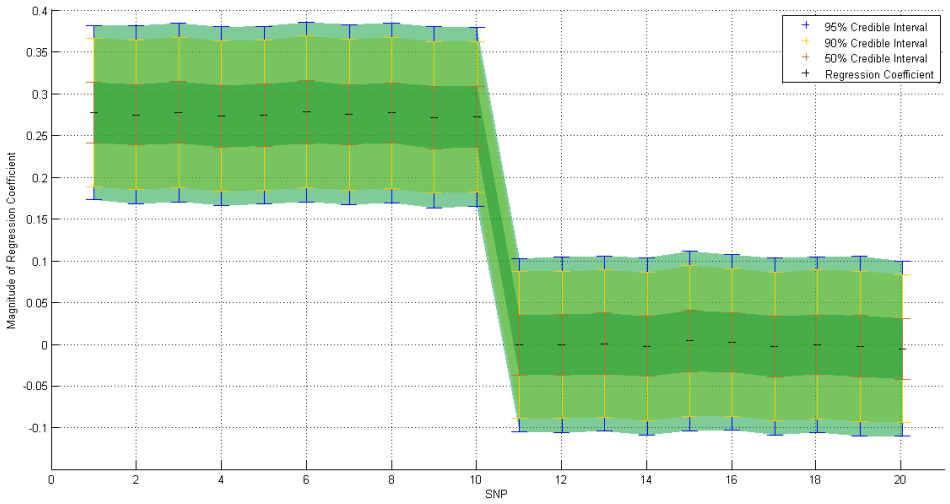

Figure 6: Mean regression coefficients and three different CIs of the 20 first predictors (SNPs) obtained by BRR when applied to one block-wise correlated data set.

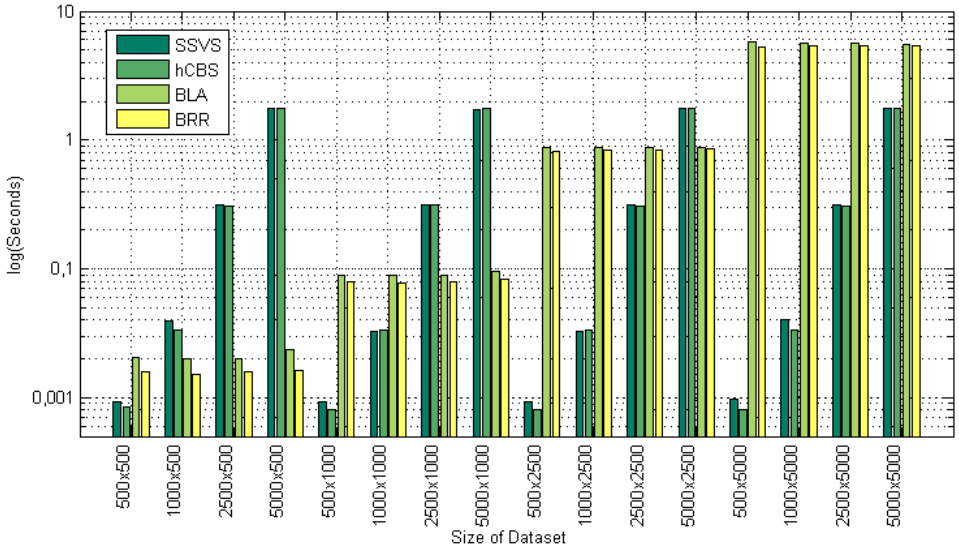

Figure 7: Computation time required for analyzing datasets of various sizes averaged over 500 iterations and 5 repeats per method.

decaying correlation. Additionally, four different scenarios, based on a real genomic dataset from cattle, were used to further assess the performance of the methods.

The main results from the investigations conducted indicate that all methods considered in this work are able to perform variable selection with a reasonable amount of true positive detections and a fair number of false positives when the information content (heritability) in datasets is high. As the information content decreases, Bayesian penalized methods are no longer able to detect associated SNPs and are outperformed by SSVS and hCBS. BLA tends to perform best in terms of highest true positives rates and low numbers of false positives in the two data sets with simulated block-wise and exponentially decreasing LD. For the data based on the two cattle chromosomes, SSVS tends to result in the highest number of true positives together with the lowest number of false positives. SSVS and hCBS mostly perform superiorly to BLA and BRR in terms of true positives, especially for the data sets with low information content. BRR detected too many false positives in most data sets.

Regarding the variable selection in BLA and BRR, using a 50% credible interval criterion did not provide useful results as the number of false positives was very large. For BLA a 95% credible interval proved to be the best choice as it increases the number of true positives with only a few more false positives. In the case of BRR a 90% credible interval also increases the number of true positives but also greatly increases the number of false positives. Therefore, using a 95% credible interval appears to be preferable.

The prediction error was in general relatively similar between all methods. However, hCBS and SSVS seems to predict unseen datasets slightly more accurately as reflected in a lower PE than the Bayesian penalized regression methods, especially for the data with block-wise correlation structure.

The computational efficiency largely depends on the size of the dataset to be analyzed as well as certain properties such as the expected number of associated SNPs in the case of the SSVS-based methods. For p≫n datasets, where the number of sequenced individuals is moderate, SSVS and hCBS are preferable. Whereas, for datasets with a similar number of SNPs and genomes sequenced, Bayesian penalized regression methods are more efficient. In more detail, Bayesian penalized regression methods scale with the number of SNPs, whereas the computational time of SSVS-based methods mostly depends on the number of phenotypes as discussed. For very large datasets with millions of SNPs, both methods may exceed computational resources.

A possible way to tackle very large datasets, which is left to be addressed in future work, would be to consider a two-step strategy where the initial selection of SNPs is performed by either SSVS or hCBS, depending on the type of dataset considered, using hyperparameters that are not overly restrictive to variable inclusion. A second step would involve computing the reduced set of SNPs using either BLA or BRR if the information content is not too low. A similar approach was used by Li et al. [11] where they first reduced the initial set of SNPs through a supervised principle component analysis and subsequently analyzed the remaining SNPs using BLA. Wilson et al. [42] used the marginal Bayes Factor (BF) to reduce the number of SNPs followed by Evolutionary Monte Carlo for Bayesian model averaging.

To our knowledge, our study is the first that evaluates different Bayesian methods on large scale GWAS data sets with p≫n based on both simulated and real LD patterns. Fridley [16] compared SSVS with Bayesian Model Averaging (BMA) and Reversible Jump MCMC (RJMCMC) on simulated and real data, but only on uncorrelated $p<n$ data. His main conclusion was that all three methods performed similar on both the simulation studies and the age-related macular degeneration (AMD) data. Rockova et al. [20] compared a range of Bayesian variable selection and regularization methods as well as frequentist methods on both simulated and real data, but also solely in the $p<n$ setting. In their simulation study, they showed that the Bayesian variable selection methods led to improved performance in detecting the true underlying model, when compared with the frequentist methods. Among the Bayesian approaches, none could be proposed as the best for all the studied simulation settings. However, one of the patterns considered by Rockova et al. [20], also addressed in our study, was that the Bayesian regularization methods appears to detect too many false positives.

Related approaches to the methods considered in this work were proposed by Hans [43], where the variable selection ability of SSVS was combined with BLA, Hans [44] who introduced a combination of SSVS and the Elastic Net, as well as Baragatti and Pommeret [45] who enhanced SSVS with the ridge regression like g-prior. A comparison between these methods and single-SNP Bayesian methods [9] on large scale GWAS data should be addressed in future studies.

Acknowledgement

Genotype data of 2122 Austrian Fleckvieh bulls were kindly provided by Zuchtdata EDV-Dienstleistungen GmbH, most of them originating from the project "Development of genomic prediction procedures for Fleckvieh cattle", funded by Österreichische Forschungsförderungsgesellschaft (FFG). The work of PW was partly funded by the Austrian Science Fund (FWF) in the framework of projcet TRP 46-B19.

References

1. Stranger BE, Stahl EA, Raj T (2011) Progress and promise of genome-wide association studies for human complex trait genetics. Genetics 187: 367-383.
2. Bush WS, Moore JH (2012) Chapter 11: Genome-wide association studies. PLoS Comput Biol 8: e1002822.
3. Balding DJ (2006) A tutorial on statistical methods for population association studies. Nat Rev Genet 7: 781-791.
4. Moore JH, Asselbergs FW, Williams SM (2010) Bioinformatics challenges for genome-wide association studies. Bioinformatics 26: 445-455.
5. Hindorff L, MacArthur J, Wise A, Jnkins H, Hall P, et al. (2013) A catalog of published genome-wide association studies. National Human Genome Research Institute.
6. Brachi B, Morris GP, Borevitz JO (2011) Genome-wide association studies in plants: the missing heritability is in the field. Genome Biol 12: 232.
7. Dekkers JC (2012) Application of genomics tools to animal breeding. Curr Genomics 13: 207-212.
8. Cherkassky V, Ma Y (2009) Another look at statistical learning theory and regularization. Neural Netw 22: 958-969.
9. Stephens M, Balding DJ (2009) Bayesian statistical methods for genetic association studies. Nat Rev Genet 10: 681-690.
10. Efron B (2010) Large-Scale Inference. Cambridge Univ Press, New York.
11. Li J, Das K, Fu G, Li R, Wu R (2011) The Bayesian lasso for genome-wide association studies. Bioinformatics 27: 516-523.
12. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, et al. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9: 356-369.
13. Szymczak S, Biernacka JM, Cordell HJ, González-Recio O, König IR, et al. (2009) Machine learning in genome-wide association studies. Genet Epidemiol 33 Suppl 1: S51-57.
14. Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B 58: 267-288.
15. Fan J, Lv J (2010) A Selective Overview of Variable Selection in High Dimensional Feature Space. Stat Sin 20: 101-148.
16. Fridley BL (2009) Bayesian variable and model selection methods for genetic association studies. Genet Epidemiol 33: 27-37.
17. O'Hara RB, Sillanpää MJ (2009) A review of Bayesian variable selection methods: What, how and which. Bayesian Analysis 4: 85-117.
18. Kyung M, Gill J, Ghosh M, Casella G (2010) Penalized regression, standard errors, and Bayesian lassos. Bayesian Analysis 5: 369-411.
19. Richardson S, Bottolo L, Rosenthal JS, Richardson S (2010) Bayesian models for sparse regression analysis of high dimensional data. In: Bayesian Statistics (9thedn), Bernardo JM, Bayarri MJ, Berger JO, Dawid AP, Heckerman D, et al., Oxford University Press, Oxford, UK.
20. Rockova V, Lesaffre E, Luime J, Löwenberg B (2012) Hierarchical Bayesian formulations for selecting variables in regression models. Stat Med 31: 1221-1237.
21. George EI, McCulloch RE (1993) Variable selection via Gibbs sampling. Journal of the American Statistical Association 88: 881.
22. Yi N, George V, Allison DB (2003) Stochastic search variable selection for identifying multiple quantitative trait loci. Genetics 164: 1129-1138.
23. Srivastava S, Chen L (2009) Comparison between the stochastic search variable selection and the least absolute shrinkage and selection operator for genome-wide association studies of rheumatoid arthritis. BMC Proc 3 Suppl 7: S21.
24. Guan Y, Stephens M (2011) Bayesian variable selection regression for genome-wide association studies and other large-scale problems. The Annals of Applied Statistics 5: 1780-1815.
25. Chen CC, Schwender H, Keith J, Nunkesser R, Mengersen K, et al. (2011)

Methods for identifying SNP interactions: a review on variations of Logic Regression, Random Forest and Bayesian logistic regression. IEEE/ACM Trans Comput Biol Bioinform 8: 1580-1591.

26. Skarman A, Shariati M, Jans L, Jiang L, Sørensen P (2012) A Bayesian variable selection procedure to rank overlapping gene sets. BMC Bioinformatics 13: 73.
27. Yi N, Xu S (2008) Bayesian LASSO for quantitative trait loci mapping. Genetics 179: 1045-1055.
28. Cai X, Huang A, Xu S (2011) Fast empirical Bayesian LASSO for multiple quantitative trait locus mapping. BMC Bioinformatics 12: 211.
29. Silva FF, Varona L, de Resende MDV, Filho JSSB, Rosa GJM, et al. (2011) A note on accuracy of Bayesian LASSO regression in GWS. Livestock Science 142: 310-314.
30. Donnelly P (2008) Progress and challenges in genome-wide association studies in humans. Nature 456: 728-731.
31. Chipman H (1996) Bayesian variable selection with related predictors. Canadian Journal of Statistics 24: 17-36.
32. Liang Y, Kelemen A (2008) Statistical advances and challenges for analyzing correlated high dimensional SNP data in genomic study for complex diseases. Statistics Surveys 2: 43-60.
33. Li F, Zhang NR (2010) Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. Journal of the American Statistical Association 105: 1202-1214.
34. Kwon D, Landi MT, Vannucci M, Issaq HJ, Prieto D, et al. (2011) An Efficient Stochastic Search for Bayesian Variable Selection with High-Dimensional Correlated Predictors. Comput Stat Data Anal 55: 2807-2818.
35. George EI, McCulloch RE (1997) Approaches for Bayesian variable selection. Statistica Sinica 7: 339-373.
36. Brown P, Vannucci M, Fearn T (1998) Bayesian wavelength selection in multicomponent analysis. Journal of Chemometrics 182: 173-182.
37. Brown P, Vannucci M, Fearn T (1998b) Multivariate Bayesian variable selection and prediction. Journal of the Royal Statistical Society: Series B 60: 627-641.
38. Brown P, Vannucci M, Fearn T (2002) Bayes model averaging with selection of regressors. Journal of the Royal Statistical Society: Series B 64: 519-536.
39. Li Q, Lin N (2010) The Bayesian elastic net. Bayesian Analysis 5: 151-170.
40. Palmer LJ, Timpson NJ, Evans DM, Davey Smith G, Cardon LR (2011) Mapping complex disease genes using linkage diseqilibrium and genome-wide scans. In: An Introduction to Genetic Epidemiology. Palmer LJ, Davey Smith G, Burton PR. The Policy Press, Bristol, UK.
41. Hastie T, Zou H (2005) Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B 67: 301-320.
42. Wilson MA, Iversen ES, Clyde MA, Schmidler SC, Schildkraut JM (2010) Bayesian Model Search and Multilevel Inference for SNP Association Studies. Ann Appl Stat 4: 1342-1364.
43. Hans C (2009) Model uncertainty and variable selection in Bayesian lasso regression. Statistics and Computing 20: 221-229.
44. Hans C (2011) Elastic net regression modeling with the orthant normal prior. Journal of the American Statistical Association 106: 1383-1393.
45. Baragatti M, Pommeret D (2012) A study of variable selection using g-prior distribution with ridge parameter. Computational Statistics and Data Analysis 56: 1920-1934.