RESEARCH

BICOSS: Bayesian Iterative Conditional Stochastic Search for GWAS

Jacob Williams[1*], Marco A. R. Ferreira[1] and Tieming Ji[2]

*Correspondence: jwilliams@vt.edu
[1]Department of Statistics, Virginia Tech, Blacksburg,24061, USA
Full list of author information is available at the end of the article

Abstract

Background: Single marker analysis (SMA) with linear mixed models for genome wide association studies (GWAS) has uncovered the contribution of genetic variants to many observed phenotypes. However, SMA has weak false discovery control. In addition, when a few variants have large effect sizes, SMA has low statistical power to detect small and medium effect sizes, leading to low recall of true causal single nucleotide polymorphisms (SNPs).

Results: We present the Bayesian Iterative Conditional Stochastic Search (BICOSS) method that controls false discovery rate and increases recall of variants with small and medium effect sizes. BICOSS iterates between a screening step and a Bayesian model selection step. A simulation study shows that, when compared to SMA, BICOSS dramatically reduces false discovery rate and allows for smaller effect sizes to be discovered. Finally, two real world applications show the utility and flexibility of BICOSS.

Conclusions: When compared to widely used SMA, BICOSS provides higher recall of true SNPs while dramatically reducing false discovery rate.

Keywords: Bayesian method; GWAS; Model Selection

Background

Genome wide association studies (GWAS) have been used successfully to identify genes involved with complex traits in a wide variety of species. To identify these genes a statistical analysis is performed to identify which single nucleotide polymorphisms (SNPs) are associated with a trait. The most common form of statistical analysis is single marker analysis (SMA) performed under the mixed model framework [1]. Algorithms such as EMMA [2] (which uses spectral decomposition of the covariance matrix for fast computation), population parameters previously determined (P3D) [3] (which speeds up computation by using the estimates of the variance components from a null model), and EMMAX [4] (which further speeds up computations of EMMA by using the estimate of the heritability from a null model) have led to widespread adoption of the mixed model framework. However, SMA has drawbacks due to not taking into account the correlation structure among SNPs, which leads to high false discovery rate (FDR) and low recall of true causal SNPs [5].

To increase recall and decrease FDR in GWAS, we propose the Bayesian Iterative Conditional Stochastic Search (BICOSS) method. Under a mixed effects model, BICOSS combines Bayesian SMA and Bayesian model selection in an iterative procedure. Each BICOSS iteration has two steps: screening and model selection. BICOSS

is initialized with the residuals from a base model that is a linear mixed model with no SNPs. Then the screening step fits as many models as the number of available SNPs, where each model has only one additional SNP and is regressed against the residuals of the base model. This screening step provides a set of candidate SNPs. The second step of BICOSS performs a model search where the possible models contain the base model and any number of SNPs from the set of candidate SNPs. When the model space is too large for complete enumeration, BICOSS performs model selection using Bayesian model selection implemented with a genetic algorithm (GA). The best model found in the model selection step becomes the base model. The next iteration of BICOSS then uses this base model to perform the screening and selection steps. BICOSS iterates between these steps until convergence of the best model. Further details as well as a graphical representation of BICOSS are provided in the Methods section. A simulation study shows that, when compared to SMA, BICOSS reduces false discovery rate and allows for SNPs with smaller effects sizes to be discovered.

Each iteration of BICOSS conditions on a base model found as the best model in the previous iteration. A key insight gained from our simulation study is that, when compared to SMA, conditioning on SNPs of high importance reduces the error variance thus allowing SNPs with smaller effect sizes to be detected. Other previous works have also used conditional models to find causal SNPs with smaller effect sizes [6, 7, 8, 9, 10]. Therefore, by conditioning on SNPs with larger effect sizes found in previous iterations, BICOSS can identify SNPs with smaller effect sizes.

A critical contribution of BICOSS is to combine model selection and screening with conditional models in an iterative procedure. This is important because model selection alone has better FDR control than single marker tests but it tends to have smaller recall. By combining the screening and model selection steps in an iterative procedure, BICOSS consistently increases recall and decreases FDR. To the best of our knowledge, there are only two other GWAS iterative procedures: GWASelect [11] and GWASinlps [12]. Both GWASinlps and GWASelect operate under the simple linear regression framework while BICOSS uses mixed effect regression. GWASelect applies SMA to a large number of bootstrap datasets followed by a LASSO procedure to identify SNPs of interest from conditional models. GWASinlps selects SNPs under a linear regression model using R^2. From the set of SNPs, GWASinlps uses Bayesian model selection with nonlocal priors to identify a best SNP model. The two main differences between BICOSS and GWASinlps are that BICOSS uses Bayesian model selection to identify candidate SNPs instead of R^2 and BICOSS uses mixed effect models instead of a linear regression models. With the publicly available code for GWASinlps, we compare GWASinlps to BICOSS in the simulation study.

Methods

BICOSS assumes the general linear mixed model ([1, 2]),

$$\mathbf{Y} = X\boldsymbol{\beta} + Z\mathbf{u} + \boldsymbol{\epsilon}, \quad \text{with} \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2 I) \quad \text{and} \quad \mathbf{u} \sim N(0, \sigma^2 \tau K), \quad (1)$$

where $\mathbf{Y}$ is a n-dimensional vector of observed phenotypes, X is an $n \times p$ matrix with columns including SNPs, intercept, and fixed effects, $\boldsymbol{\beta}$ is a p-dimensional vector

of regression coefficients, Z is an $n \times t$ incidence matrix mapping each observed phenotype to one of t inbred strains, $\mathbf{u}$ is a t-dimensional vector of random effects accounting for population structure, and ϵ is an error term. In addition, σ^2 is the variance of the unstructured error and τ a kinship dependence parameter. Finally, K is the realized relationship matrix or kinship matrix assumed to be a known positive semi-definite matrix.

Figure 1 presents a graphical representation of BICOSS. BICOSS is an iterative procedure where each iteration is comprised of two steps: a screening and a model selection step. BICOSS is initialized with a base model fitted as a linear mixed model with no SNPs in the model. Then the screening step fits as many models as there are SNPs, each model containing one SNP and regressed against the residuals of the base model. The screening step identifies a set of candidate SNPs using Bayesian FDR control applied to the posterior probabilities of the SNPs. Then, the model selection step of BICOSS performs Bayesian model selection where the possible models contain any combination of the base model and SNPs from the candidate set. If the model space is too large to perform complete enumeration, a genetic algorithm is used to perform stochastic model search. The model with the highest posterior probability is the best model. This best model becomes the base model for the next iteration which proceeds with the screening and model selection steps. BICOSS iterates these two steps until convergence of the best model.

Flowchart1.PNG

Figure 1 Graphical Representation of BICOSS.

We cast both the screening step and model selection step within a Bayesian model selection framework. We briefly highlight Bayesian model selection and the priors on the model space before providing the full derivation for the screening and model selection steps.

Bayesian Model Selection

Bayesian model selection assumes m possible models $M_1, \ldots, M_m$. Let $P(M_i)$ be the prior model probability for model M_i. In addition, assume that the unknown parameters in model M_i are collected in parameter vector $\theta_i \in \Theta_i$ and have prior density $\pi(\theta_i)$. Let the dimension of θ_i be d_i. Finally, assume the likelihood function under model M_i is $\mathcal{L}(\mathbf{Y} \mid \theta_i, M_i)$. Thus, an important quantity in Bayesian model selection is the marginal likelihood under model M_i, $i = 1, \ldots, n$, given by

$$m_i(\mathbf{Y}) = \int_{\Theta_i} \mathcal{L}(\mathbf{Y} \mid \theta_i, M_i)\pi(\theta_i)d\theta_i. \tag{2}$$

Hence, by Bayes Theorem the posterior probability of model M_i given the data $\mathbf{Y}$ is

$$P(M_i|\mathbf{Y}) = \frac{P(M_i)m_i(\mathbf{Y})}{\sum_{j=1}^{m} P(M_j)m_j(\mathbf{Y})}. \tag{3}$$

Assuming a base model M_b, the Bayes factor of model M_i with respect to M_b is defined as $BF_{ib} = m_i(\mathbf{Y})/m_b(\mathbf{Y})$. Hence, the posterior probability of model M_i given the data $\mathbf{Y}$ can be computed as

$$P(M_i|\mathbf{Y}) = \frac{P(M_i)BF_{ib}}{\sum_{j=1}^{m} P(M_j)BF_{jb}}. \tag{4}$$

Now, let the BIC of model M_i be

$$BIC_i = -2\log\left(\mathcal{L}(\mathbf{Y} \mid \hat{\theta}_i, M_i)\right) + d_i \log(n), \tag{5}$$

where $\hat{\theta}_i$ is the maximum likelihood estimate of θ_i. The Bayes factor BF_{ib} can be approximated by using the BIC ([13, 14]). Specifically, if the information contained in each prior $\pi(\theta_i)$ is equivalent to one observation, then the Bayes factor BF_{ib} can be approximated with

$$BF_{ib} \approx \exp\left\{-0.5(BIC_i - BIC_b)\right\}, \tag{6}$$

with error $\mathcal{O}(n^{-1/2})$ [15]. With this approximation, we do not need to explicitly specify the prior densities $\pi(\theta_i)$. BICOSS uses this approximation combined with Eqn. 4 to compute the posterior probabilities of the competing models.

Prior Model Probabilities in BICOSS

Consider a model with s possible SNPs. Following standard practice in modern Bayesian model selection, we treat the inclusion of each of the possible s SNPs as independent Bernoulli trials with success probability $(1-\pi_0)$. As a result, the prior probability of model M_i is

$$P(M_i) = (\pi_0)^{s-p_i}(1-\pi_0)^{p_i}, \tag{7}$$

where p_i is the number of SNPs in model M_i. Here, we estimate the true rate of null hypothesis π_0 using the procedure proposed in [16] which uses the p-values of each SNP from a SMA to calculate the estimated proportion of true null SNPs. When this procedure conservatively estimates $\pi_0 = 1$, BICOSS sets $\pi_0 = 1 - 100L^{-1}$ where L is the total number of SNPs. The p-values are calculated at every screening step, therefore the estimate of π_0 is updated at every iteration of the screening step of BICOSS. The model selection step uses the same π_0 estimated at the first screening, which allows SNPs that were detected in the first screening to be competitive in the model selection compared to SNPs found in subsequent iterations.

Screening Step

The screening step starts by fitting the base model which is obtained from Eqn. 1 with the matrix X containing the SNPs from the base model of the previous iteration of BICOSS. From this base model fit, we obtain estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\tau}$. Let $\hat{\mathbf{Y}} = Y - X\hat{\boldsymbol{\beta}}$ and let $\Sigma(\hat{\tau}) = (I-P)(I+\hat{\tau}K)(I-P)$ where $P = X(X^\top(I+\hat{\tau}K)^{-1}X)^{-1}X^\top$ is a

projection matrix. Recall that L is the total number of SNPs. Then the screening step fits for each SNP l, $l = 1, \ldots, L$, the linear mixed model

$$\hat{\mathbf{Y}} = X_l \beta_l + \boldsymbol{\epsilon}^*, \quad \boldsymbol{\epsilon}^* \sim N(\mathbf{0}, \sigma^2 \Sigma(\hat{\tau})), \tag{8}$$

where X_l is an $n \times 1$ vector for SNP l.

In the screening step, for each SNP l we compare only two models: the base model, and the base model with the added SNP l. In that context, Eqn. 4 in the section on Bayesian model selection is used to compute the posterior probability of SNP l being a causal SNP conditional on the base model. The screening step then scans through all SNPs computing these posterior probabilities.

To control the false discovery rate, BICOSS uses Bayesian FDR control ([17, 18, 19, 20]). Let $r_l = 1$ if SNP l is a true causal SNP and $r_l = 0$ otherwise. Let $p_l = P(r_l = 1|\mathbf{Y})$ which is computed as described in the above paragraph using the Bayes factor comparing the model with SNP l versus the model without SNP l. Then a possible decision rule is to flag SNP l as significant if p_l is greater than or equal to a threshold p_0. The resulting FDR is then equal to

$$FDR = \frac{\sum_{l=1}^{L}(1 - r_l)1_{p_l \geq p_0}}{\sum_{l=1}^{L} 1_{p_l \geq p_0}}, \tag{9}$$

where 1 denotes the indicator function. Further, because the true value of r_l is unknown the posterior expected value of the FDR given the data can be estimated as

$$\widehat{FDR} = \frac{\sum_{l=1}^{L}(1 - p_l)1_{p_l \geq p_0}}{\sum_{l=1}^{L} 1_{p_l \geq p_0}}. \tag{10}$$

A more desired decision rule would be to control for the desired nominal FDR level denoted as q_0 rather than an arbitrary predetermined threshold p_0. Specifically, we first rank the SNPs in decreasing order of p_l. Denote the ordered estimates of the posterior model probabilities as $\{p_{(1)}, p_{(2)}, \ldots, p_{(L)}\}$. Thus, denoting $d \in \{1, \ldots, L\}$, the posterior expected FDR of selecting the first d ordered SNPs as significant is

$$\widehat{FDR}_d = \frac{\sum_{l=1}^{L}(1 - p_l)1_{p_l > p_{(d)}}}{\sum_{l=1}^{L} 1_{p_l > p_{(d)}}} = \frac{\sum_{l=1}^{d}(1 - p_{(l)})}{d}. \tag{11}$$

The decision rule for detecting causal SNPs is to flag all SNPs with $\widehat{FDR}_d < q_0$. This provides a list of candidate SNPs for the BICOSS selection step. The simulation study and the real data analyses use $q_0 = 0.05$.

Model Selection Step

With the list of candidate SNPs from the screening step, the model selection step performs a model search where the possible models include any combination of SNPs in the base model and the candidate SNPs identified in the latest screening. Each possible model is evaluated using the Bayesian model selection procedure described in the section on Bayesian model selection with prior model probability

given in Eqn. 7. To accelerate computation, we take a P3D approach and estimate the kinship dependence parameter τ only once based on the full model that includes the SNPs from the base model as well as the candidate SNPs. This parameter τ is kept fixed at this estimate when fitting all other models.

Depending on the number of SNPs identified in the screening step, one of two different algorithms are used to search the model space. When the dimensionality is low, a complete enumeration is used to compute posterior model probabilities for every possible model. When the number of SNPs is high such that complete enumeration would be computationally expensive (16 or more), a genetic algorithm is used to search the model space.

BICOSS uses a genetic algorithm implemented in the R package GA [21] that iterates mutation, crossover, and selection steps. The genetic algorithm starts with a population of 100 models. One of these models has just the intercept. Another set of models in this initial population has only one SNP per model, where the SNPs are either from the base model or are candidate SNPs. If there are more than 99 of these SNPs, then the 99 SNPs with the highest posterior probabilities are used to initialize the initial population. If there are less than 99 of these SNPs, then the remaining models in the initial population are chosen at random. The mutation, crossover, and selection steps then operate on the population to create subsequent populations. The mutation step creates a new model from an existing model by changing the status of a SNP in that model, e.g. if a SNP is present in the existing model it will become absent in the new model. The crossover step creates two models by combining two existing models. Finally, the selection step samples models to be passed to the next population with probabilities proportional to $\exp(-0.5BIC_i)$ for model M_i.

We consider two different convergence criteria, 400 maximum iterations or 40 consecutive iterations with the same best model, whatever happens first. We also considered convergence criteria with 4,000 maximum iterations and 400 iterations with the same best model, but the results were about the same. We report the results for the latter set of convergence criteria in the supplementary material. If the best model identified in the selection step matches the current base model, BICOSS converges. Otherwise, the base model is updated to be the best model found, and another iteration of BICOSS is performed.

Results

Simulation Study

We have performed a simulation study to compare BICOSS to other competing methods. In addition, we present two smaller simulation studies to evaluate the robustness of BICOSS, when there are no causal SNPs and when there is no kinship dependence structure. For all three simulation studies, we compare BICOSS to SMA methods with the Bonferroni correction and GWASinlps. We consider two SMA methods based on the linear mixed model from Eqn. 1: a method we call SMA-Exact that similarly to EMMA uses the spectral decomposition of the kinship dependence structure; and a method we call SMA-Approx. that similarly to EMMAX fixes the variance parameters at their estimates for a model without SNPs. For direct comparison of computation time, all methods are implemented in R. Both BICOSS

and SMA methods use a FDR nominal level of 0.05. The genotype data used for all three simulation studies is from 328 *A. Thaliana* accessions from the TAIR9 sequence [22]. In this simulation study $n = 328$ and $Z = I_{n\times n}$. Specifically, we consider a set of 60,000 SNPs. To obtain these 60,000 SNPs, we obtained 10 blocks of 6,000 SNPs each with minor allele frequency above 0.01 from *A. Thaliana*, where each block was separated from the subsequent block by 15,000 SNPs. Figure S1 in the supplementary material presents a heatmap of the correlation matrix of the first block with 6,000 SNPs for the 328 *A. Thaliana* accessions. For the general simulation study and the case when there is no kinship dependence structure, we placed the causal SNPs in positions 3,000, 9,000, 15,000, 21,000, 27,000, 33,000, 39,000, 45,000, 51,000, and 57,000 of the 60,000 SNPs. The kinship matrix used in the case of no causal SNPs and the general simulation study was built from the entire TAIR9 SNP array for the 328 ascensions of *A. Thaliana* using the function A.mat from the R package rrBLUP [23].

We compare the competing methods with four different criteria: recall, also known as true positive rate, FDR, False positive rate, and the F1 score. We also report computation time. Recall is defined as the number of identified true causal SNPs divided by the total number of causal SNPs. The FDR is defined as the number of false positives identified as significant divided by the number of SNPs identified as significant. The false positive rate is the number of false positives divided by the number of false positives plus the number of true negatives. The F1 score is the number of true positives divided by the number of true positives plus half the sum of false positives and false negatives. We report the computation time in seconds for each procedure using 12 cores of a 2 $\times$ 12 core Intel Xeon 2.5 GHz 12-core with 256 GB of memory running OpenBlas for optimized matrix algebra. The results presented here are for GWASinlps version 2.0 with tuning parameters $k_0 = 1$, $n_{skip} = 3$, $r_{xx} = 0.2$, $m = 500$, and $\tau = 0.022$ as recommended in both the GWASinlps documentation and in [12]. For accurate comparison of methods, the results for each simulation setting are based on 100 simulated datasets.

General Simulation Study

A general simulation study to compare BICOSS to other competing methods is conducted under the linear mixed model:

$$\mathbf{Y} = \alpha\mathbf{1} + X\boldsymbol{\beta} + Z\mathbf{u} + \boldsymbol{\epsilon}, \tag{12}$$

where $\mathbf{u} \sim N(0, \sigma^2\tau K)$, $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$, and α is 1.

We consider 10 causal SNPs with six different settings of β vectors. Seven coefficients remained fixed at 0.4 while the other three coefficients were equal to each other and assumed values of 0.05, 0.1, 0.2, 0.4, 0.8, and 1.6. Thus, the fourth setting had equal coefficients across the entire set of causal SNPs. For every simulated $\mathbf{Y}$, the values of τ and σ^2 were equal to 0.1 and 0.2 respectively, which are similar to the estimates of τ and σ^2 obtained in the case study on salt stress in *A. Thaliana*.

Table 1 displays results averaged over the 100 datasets under each setting. SMA procedures typically discover about 3 of the 10 true causal SNPs, BICOSS typically discovers about 5 to 7 causal SNPs. Therefore, while SMA methods typically discover only the SNPs with large effect sizes, BICOSS is able to discover SNPs with

Table 1 Results of simulation study with linear mixed model. Regression coefficients of causal SNPs $\boldsymbol{\beta} = (\beta^{(1)}, 0.4, 0.4, 0.4, \beta^{(1)}, 0.4, 0.4, 0.4, \beta^{(1)}, 0.4)^\top$. Average Performance of each method over 100 datasets for each setting. Recall indicates the True Positive Rate, FDR is the False Discovery Rate, FPR is the False Positive Rate, and F1 is the F1 score.

Setting	Measure	Method			
		SMA - Exact	SMA - Approx.	BICOSS	GWASinlps
Setting 1 $\beta^{(1)} = 0.05$	Recall	0.36	0.35	0.49	0.55
	FDR	0.61	0.60	0.27	0.62
	FPR $\times 10^5$	12.70	12.30	3.95	17.44
	F1	0.35	0.35	0.57	0.44
	Time (s)	197	2	22	85
Setting 2 $\beta^{(1)} = 0.1$	Recall	0.33	0.33	0.49	0.54
	FDR	0.57	0.56	0.28	0.61
	FPR $\times 10^5$	11.22	10.90	4.10	16.45
	F1	0.35	0.35	0.57	0.44
	Time (s)	203	2	46	122
Setting 3 $\beta^{(1)} = 0.2$	Recall	0.31	0.31	0.49	0.55
	FDR	0.61	0.61	0.34	0.63
	FPR $\times 10^5$	11.17	10.87	5.02	19.02
	F1	0.33	0.33	0.55	0.42
	Time (s)	201	2	47	117
Setting 4 $\beta^{(1)} = 0.4$	Recall	0.34	0.33	0.58	0.65
	FDR	0.59	0.58	0.34	0.62
	FPR $\times 10^5$	10.47	10.22	5.50	20.14
	F1	0.35	0.35	0.61	0.47
	Time (s)	203	2	50	130
Setting 5 $\beta^{(1)} = 0.8$	Recall	0.29	0.28	0.73	0.79
	FDR	0.79	0.79	0.33	0.60
	FPR $\times 10^5$	21.60	21.35	6.93	22.25
	F1	0.23	0.23	0.69	0.52
	Time (s)	186	1	44	148
Setting 6 $\beta^{(1)} = 1.6$	Recall	0.30	0.30	0.70	0.78
	FDR	0.92	0.92	0.30	0.65
	FPR $\times 10^5$	61.23	60.49	5.70	25.99
	F1	0.12	0.12	0.69	0.48
	Time (s)	176	1	45	147

Table 2 Results of simulation study with no causal SNPs. Average Performance of each method over 100 datasets. FP indicates the number of false positives.

Setting	Measure	Method			
		SMA - Exact	SMA - Approx.	BICOSS	GWASinlps
No Causal SNPs	FP	0.05	0.04	1.33	8.13
	Time (s)	197	2	22	85

smaller effect sizes. In addition, BICOSS maintains a substantially lower FDR, lower FPR, and higher F1 score in all settings compared to SMA. The massive improvement in these measures is due to the model selection step. Specifically, by allowing multiple SNPs to compete in the best model, BICOSS model selection step better controls FDR.

Compared to GWASinlps, BICOSS provides a similar recall while yielding a much lower FDR, lower FPR, and higher F1 score. BICOSS is more conservative overall than GWASinlps, but the F1 score (that is, the harmonic mean of precision and recall) highlights the improved combined performance in terms of recall and FDR of BICOSS compared to GWASinlps. The better performance of BICOSS when compared to GWASinlps may be explained by two main reasons. First, BICOSS uses a Bayesian screening step while GWASinlps uses a R^2-based screening. Second, BICOSS assumes a linear mixed model whereas GWASinlps assumes a linear model with independent errors. In particular, the linear mixed model assumed by BICOSS is more realistic in the context of GWAS analysis.

Our simulation study also shows that when some few SNPs have very large effect sizes as in Settings 5 and 6, SMA methods have difficulty identifying SNPs with medium effect sizes and produce very large FDR. Specifically, Table 1 shows that, in Settings 5 and 6, SMA methods can only find 30% of the true causal SNPs and has FDR of 0.79 and 0.92 respectively. In contrast, in these settings BICOSS has recall at or above 70% and much better FDR control.

Robustness to Lack of Signal

To examine the robustness of BICOSS when applied to datasets with no causal SNPs, we simulate 100 datasets from the model:

$$\mathbf{Y} = \alpha\mathbf{1} + Z\mathbf{u} + \boldsymbol{\epsilon}, \tag{13}$$

where $\mathbf{u} \sim N(0, \sigma^2\tau K)$, $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$, and $\alpha = 1$. Similarly, for every simulated $\mathbf{Y}$, the values of τ and σ^2 were equal to 0.1 and 0.2 respectively, which are similar to the estimates of τ and σ^2 obtained in the case study on salt stress in *A. Thaliana*. As there are no true causal SNPs in Equation 13, we only examine the number of false positives.

Table 2 presents the results for the 100 simulated datasets under this scenario. In this case, SMA methods have a stricter control of false positives compared to the two iterative procedures. BICOSS performs significantly better than GWASinlps but is not as conservative as SMA. Therefore, one limitation of BICOSS is that it has on average a slightly larger number of false positives then SMA when applied to datasets with no causal SNPs.

Robustness to Lack of Kinship Dependence Structure

To check how BICOSS performs when the data are from a linear model without kinship dependence, we simulated 100 datasets from the linear model:

$$\mathbf{Y} = \alpha\mathbf{1} + X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{14}$$

where $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$, $\alpha = 1$, and $\sigma^2 = 0.2$. Note that BICOSS has been built using the mixed model framework. Meanwhile, GWASinlps was built assuming a linear model. Thus, in principle, data simulated from Equation 14 should favor GWASinlps. We explore one setting of $\boldsymbol{\beta}$, all causal coefficients equal to 0.4. Thus, this simulation has identical X and $\boldsymbol{\beta}$ as setting 4 of the general simulation. P-values are calculated for SMA using the classic T statistic for simple linear regression models. Therefore as this is an exact procedure we show results labeled as SMA-Exact.

Table 3 presents the results of the linear model simulation study. Similar to the simulation with linear mixed models, BICOSS has the lowest FDR, lowest FPR, and highest F1. This is not completely surprising because for datasets simulated from Equation 14, the kinship dependence parameter τ is usually estimated as very small. In the limit when τ is estimated to be 0, the linear mixed model in Equation 1 becomes a linear model. Therefore, even when there is no kinship structure, BICOSS is able to automatically adapt and perform better than competing methods.

Case Studies

To demonstrate the utility and flexibility of BICOSS, we present two case studies with real data analyses. First, BICOSS is implemented on data from a published study of salt stress on the selfing species *A. Thaliana* [24]. Second, BICOSS is applied to a study of alcohol dependency in humans.

Salt Stress in A. Thaliana

This study considers three different settings of soil salt stress to evaluate which genes are potentially impactful [24]. The three settings considered were a control setting, 75 mM of NaCl, and 125 mM of NaCl. Different measures of the root structure were taken to gauge how salt stress impacted the plants. In this case study, we analyze the average length of lateral root per main root length for 328 *A. Thaliana* accessions under 75 mM NaCl salt stress. Genotype data was obtained from TAIR9 [22]. Only SNPs with minor allele frequency greater than 0.01 were included, thus the analysis presented here considers approximately 213,000 SNPs.

Table 3 Results of simulation study with linear model. Regression coefficients of causal SNPs $\boldsymbol{\beta} = (0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4)^\top$. Average Performance of each method over 100 datasets for each setting. Recall indicates the True Positive Rate, FDR is the False Discovery Rate, FPR is the False Positive Rate, and F1 is the F1 score.

Setting	Measure	Method		
		SMA - Exact	BICOSS	GWASinlps
Linear Model	Recall	0.38	0.61	0.66
	FDR	0.62	0.38	0.62
	FPR $\times 10^5$	12.00	6.95	20.34
	F1	0.37	0.60	0.47
	Time (s)	6	55	169

Table 4 The number of SNPs identified by method for each case study. Multiple comparison corrections use nominal level 0.05 and are based on the number of SNPs in a given genotype dataset.

Method	Salt Stress in A. Thaliana		AUD in Humans	
	Number of SNPs	Time (s)	Number of SNPs	Time (m)
SMA - Exact	22	555	15	82
SMA - Approx.	22	8	15	4
BICOSS	5	142	6	38
GWASinlps	37	544	499	792

Table 4 presents the number of SNPs found by SMA, BICOSS, and GWASinlps as well as the computational time. For *A. Thaliana*, both SMA methods found 22 SNPs, GWASinlps found 37 SNPs, and BICOSS identified just 5 SNPs. Similar to the simulation study, we see a large difference in the total number of SNPs found by BICOSS when compared to SMA and GWASinlps. Surprisingly, we note a large increase of the total number of SNPs found by GWASinlps compared to SMA. Given the results of the simulation study, we expect the majority of SNPs found by GWASinlps and SMA methods to be false positives. Based on the simulation study, BICOSS has a much better control of FDR than the other methods. Thus, for purpose of discussion we will focus on the results from BICOSS. Of the five SNPs identified by BICOSS, one SNP is perfectly correlated to two other SNPs, implying seven identified SNPs.

The seven SNPs are in genes AT1G62500, AT2G38970, AT3G60370, AT4G14305, AT4G39955, AT4G39970, and AT4G40000. Previous literature relates two of these genes to response to salt stress. Specifically, AT1G62500 is a differentially expressed gene which has been shown to activate in the event of salt stress [25]. In addition; AT4G39955 is an α/β-Hydrolases superfamily protein. α/β-Hydrolases superfamily proteins have been shown to enhance salt tolerance in the sweet potato family [26].

Alcohol Use Disorder in Humans

In this case study, we use publicly available data from The Collaborative Study on the Genetics of Alcoholism (COGA) that was performed to identify novel genetic factors associated with alcohol use disorder (AUD) [27]. Specifically, in this case study we analyze the response variable "age of first drink", for 1738 people of European ancestry with approximately 1 million sequenced SNPs. To normalize and variance-stabilize the data, the logarithm transformation was applied to age of first drink. Only SNPs with minor allele frequency larger than 0.01 were investigated for this analysis. Further, any SNP that did not have an rsID or was located in chromosome X or Y was removed from the analysis. Thus, this analysis considers approximately 840,000 SNPs.

Table 4 presents the number of SNPs found by SMA, BICOSS and GWASinlps and the timing of each method. Similarly to the simulation study and the *A. Thaliana* case study, SMA and GWASinlps identified large numbers of SNPs. Specifically for the AUD case study, both SMA methods found 15 SNPs, GWASinlps found 499 SNPs, and BICOSS found just 6 SNPs. Because BICOSS has a much better FDR control than the other methods, here we investigate the genes found by BICOSS. BICOSS identified six SNPs, which are in the following genes: KCNMA1, ZYG11A, TPTE2, ABCF1, ANKS1B, and LINC02237. LINC02237 is a long intergenic non-protein coding RNA and the other genes are all protein coding genes.

Of the five protein coding genes found by BICOSS, two have published associations with AUD and two have been linked to liver diseases. Specifically, KCNMA1 is known as a gene associated with alcohol dependency [28]. In addition, in a study with people of Chinese Han ethnicity, ANKS1B has been found to be associated with alcoholism [29]. Further, TPTE2 has been shown to be related to hepatic fibrogenesis and fibrosis [30]; alchohol abuse is one of the main causes of liver fibrosis [31]. Furthermore, ABCF1 has been shown to be overexpressed in hepatocellular carcinoma [32]. These results indicate possibly important genes for further potential investigation for a better understanding of alcohol use disorder.

Discussion

We have presented BICOSS, a novel Bayesian method for the analysis of GWAS data. To take into account the correlation structure among SNPs, BICOSS iterates a screening step and a model selection step. Simulation studies show that, while when there are no true SNPs BICOSS tends to identify a slightly larger number of SNPs than SMA methods, when there are true causal SNPs, BICOSS performs much better than SMA. In the latter case when compared to SMA, BICOSS has greater recall of true causal SNPs while maintaining a much lower FDR. In addition, when there are SNPs with large effect sizes, BICOSS has increased recall of true causal SNPs with small and medium effect sizes. Further, when compared to the Bayesian iterative method GWASinlps, BICOSS maintains comparable recall while having a much lower FDR.

While here we have implemented BICOSS within the EMMAX [4] methodology, we note that BICOSS can be easily adapted to work with other GWAS frameworks such as GCTA [33]. Applying BICOSS should be relatively straightforward when the model and the likelihood can be explicitly written.

There are many possible avenues for future research. For example, a potentially useful avenue is to extend BICOSS to use explicit prior distributions for the parameters. Such extension would allow the incorporation of substantive prior information in the GWAS analysis. Another possible area of research would be to extend BICOSS to BioBank scale data. Finally, another possible area of research would be to extend BICOSS for the analysis of non-Gaussian data such as the number of lateral roots in *A. Thaliana* or the indicator of alcohol dependency for families with members suffering alcohol use disorder.

Conclusion

We propose BICOSS, a novel iterative Bayesian procedure for GWAS analysis. Compared to SMA, BICOSS increases recall of true causal SNPs while dramatically reducing FDR. Upon publication of this article, BICOSS will be made available in the R package GWAS.BAYES that is available of Bioconductor.

Funding

This work was supported by National Science Foundation grants DMS 1853549, DMS 1853556, and DMS 2054173.

Abbreviations

Single marker analysis (SMA),Genome-Wide Association Studies (GWAS), Single Nucleotide Polymorphisms (SNPs), Bayesian Iterative Conditional Stochastic Search (BICOSS), False Discovery Rate (FDR), Genetic Algorithm (GA), Bayesian Information Criterion (BIC), Minor Allele Frequency (MAF).

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and materials
The two case study datasets are publicly available from the following websites. A. Thaliana phenotype: https://arapheno.1001genomes.org; A. Thaliana genotype: dataset available from R package qtcat.data (https://rdrr.io/github/QTCAT/qtcat.data/; and genotype and phenotype data for Alcohol use disorder in humans: www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/collection.cgi?study_id=phs000092.v1.p1.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
JW, MARF, and TJ conceived the study. JW and MARF developed the methodology and simulation experiments. JW implemented the simulation experiments. JW implemented the methodology and analyzed the results supervised by MARF. MARF and TJ acquired the funding. JW and MARF wrote the manuscript. JW, MARF, and TJ reviewed the manuscript. All authors read and approved the final manuscript.

Acknowledgements
This version of the article has been accepted for publication, after peer review but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: https://doi.org/10.1186/s12859-022-05030-0. Computations for this manuscript have been performed on supercomputers of Advanced Research Computing at Virginia Tech.

Author details
[1]Department of Statistics, Virginia Tech, Blacksburg,24061, USA. [2] Biostatistics, GRAIL, 94025, Menlo Park, USA.

References
1. Yu, J., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., *et al.*: A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics **38**(2), 203–208 (2006)
2. Kang, H.M., Zaitlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., Daly, M.J., Eskin, E.: Efficient Control of Population Structure in Model Organism Association Mapping. Genetics **178**(3), 1709–1723 (2008)
3. Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R.J., Tiwari, H.K., Gore, M.A., Bradbury, P.J., Yu, J., Arnett, D.K., Ordovas, J.M., *et al.*: Mixed linear model approach adapted for genome-wide association studies. Nature genetics **42**(4), 355–360 (2010)
4. Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.-y., Freimer, N.B., Sabatti, C., Eskin, E., *et al.*: Variance component model to account for sample structure in genome-wide association studies. Nature Genetics **42**(4), 348–354 (2010)
5. Stringer, S., Wray, N.R., Kahn, R.S., Derks, E.M.: Underestimated effect sizes in gwas: Fundamental limitations of single snp analysis for dichotomous phenotypes. PloS one **6**(11), 27964 (2011)
6. Yang, J., Ferreira, T., Morris, A.P., Medland, S.E., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., Weedon, M.N., Loos, R.J., *et al.*: Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nature Genetics **44**(4), 369–375 (2012)
7. Dolejsi, E., Bodenstorfer, B., Frommlet, F.: Analyzing genome-wide association studies with an FDR controlling modification of the Bayesian Information Criterion. PIOS ONE **9**(7), 103322 (2014)
8. Allen, H.L., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., *et al.*: Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature **467**(7317), 832–838 (2010)
9. Ripke, S., Sanders, A.R., Kendler, K.S., Levinson, D.F., Sklar, P., Holmans, P.A., Lin, D.-Y., Duan, J., Ophoff, R.A., Andreassen, O.A., *et al.*: Genome-wide association study identifies five new schizophrenia loci. Nature Genetics **43**(10), 969 (2011)
10. Sklar, P., Ripke, S., Scott, L.J., Andreassen, O.A., Cichon, S., Craddock, N., Edenberg, H.J., Nurnberger Jr, J.I., Rietschel, M., Blackwood, D., *et al.*: Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nature Genetics **43**(10), 977 (2011)
11. He, Q., Lin, D.-Y.: A variable selection method for genome-wide association studies. Bioinformatics **27**(1), 1–8 (2011)
12. Sanyal, N., Lo, M.-T., Kauppi, K., Djurovic, S., Andreassen, O.A., Johnson, V.E., Chen, C.-H.: GWASinlps: non-local prior based iterative SNP selection tool for genome-wide association studies. Bioinformatics **35**(1), 1–11 (2019)
13. Frommlet, F., Ruhaltinger, F., Twaróg, P., Bogdan, M.: Modified versions of Bayesian Information Criterion for genome-wide association studies. Computational Statistics & Data Analysis **56**(5), 1038–1051 (2012)
14. Schwarz, G.: Estimating the dimension of a model. Annals of Statistics **6**(2), 461–464 (1978)
15. Kass, R.E., Wasserman, L.: A Reference Bayesian Test for Nested Hypotheses and its Relationship to the Schwarz Criterion. Journal of the American Statistical Association **90**(431), 928–934 (1995)
16. Langaas, M., Lindqvist, B.H., Ferkingstad, E.: Estimating the Proportion of True Null Hypotheses, with Application to DNA Microarray Data. Journal of the Royal Statistical Society. Series B (Statistical Methodology) **67**(4), 555–572 (2005)
17. Newton, M.A., Noueiry, A., Sarkar, D., Ahlquist, P.: Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics **5**(2), 155–176 (2004)

18. Müller, P., Parmigiani, G., Rice, K.: FDR and Bayesian multiple comparisons rules. In: Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A., Heckerman, D., Smith, A.F.M., West, M. (eds.) Bayesian Statistics 8, pp. 349–370. Oxford Univ. Press, Oxford (2007)
19. Cui, S., Guha, S., Ferreira, M.A.R., Tegge, A.N.: hmmseq: A hidden Markov model for detecting differentially expressed genes from RNA-seq data. The Annals of Applied Statistics **9**(2), 901–925 (2015)
20. Xie, J., Ji, T., Ferreira, M.A.R., Li, Y., Patel, B.N., Rivera, R.M.: Modeling allele-specific expression at the gene and SNP levels simultaneously by a Bayesian logistic mixed regression model. BMC Bioinformatics **20**(1), 1–13 (2019)
21. Scrucca, L.: GA: A Package for Genetic Algorithms in R. Journal of Statistical Software, Articles **53**(4), 1–37 (2013)
22. Horton, M.W., Hancock, A.M., Huang, Y.S., Toomajian, C., Atwell, S., Auton, A., Muliyati, N.W., Platt, A., Sperone, F.G., Vilhjálmsson, B.J., *et al.*: Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. Nature Genetics **44**(2), 212–216 (2012)
23. Endelman, J.B.: Ridge Regression and Other Kernels for Genomic Selection with R Package rrblup. The Plant Genome **4**(3), 250–255 (2011)
24. Julkowska, M.M., Koevoets, I.T., Mol, S., Hoefsloot, H., Feron, R., Tester, M.A., Keurentjes, J.J.B., Korte, A., Haring, M.A., de Boer, G.-J., Testerink, C.: Genetic Components of Root Architecture Remodeling in Response to Salt Stress. The Plant Cell **29**(12), 3198–3213 (2017). doi:10.1105/tpc.16.00680
25. Jing, Y., Shi, L., Li, X., Zheng, H., Gao, J., Wang, M., He, L., Zhang, W.: OXS2 is required for salt tolerance mainly through associating with salt Inducible genes, CA1 and Araport11, in Arabidopsis. Scientific Reports **9**(1), 1–11 (2019)
26. Liu, D., Wang, L., Zhai, H., Song, X., He, S., Liu, Q.: A novel α/β-hydrolase gene IbMas enhances salt tolerance in transgenic sweetpotato. PloS One **9**(12), 115128 (2014)
27. Begleiter, H., Reich, T., Hesselbrock, V., Porjesz, B., Li, T.-K., Schuckit, M.A., Edenberg, H.J., Rice, J.P., *et al.*: The Collaborative Study on the Genetics of Alcoholism. Alcohol Health and Research World **19**, 228–228 (1995)
28. Bettinger, J.C., Davies, A.G.: The role of the BK channel in ethanol response behaviors: evidence from model organism and human studies. Frontiers in Physiology **5**, 346 (2014)
29. Sun, Y., Chang, S., Liu, Z., Zhang, L., Wang, F., Yue, W., Sun, H., Ni, Z., Chang, X., Zhang, Y., *et al.*: Identification of novel risk loci with shared effects on alcoholism, heroin, and methamphetamine dependence. Molecular Psychiatry **26**(4), 1152–1161 (2021)
30. Liu, Z., Chalasani, N., Lin, J., Gawrieh, S., He, Y., Tseng, Y.J., Liu, W.: Integrative omics analysis identifies macrophage migration inhibitory factor signaling pathways underlying human hepatic fibrogenesis and fibrosis. Journal of bio-X research **2**(01), 16–24 (2019)
31. Hernandez-Gea, V., Friedman, S.L.: Pathogenesis of liver fibrosis. Annual review of pathology: mechanisms of disease **6**, 425–456 (2011)
32. Fung, S.W., Cheung, P.F.-Y., Yip, C.W., Ng, L.W.-C., Cheung, T.T., Chong, C.C.-N., Lee, C., Bo-San Lai, P., Chan, A.W.-H., Tsao, G.S.-W., *et al.*: The atp-binding cassette transporter abcf1 is a hepatic oncofetal protein that promotes chemoresistance, emt and cancer stemness in hepatocellular carcinoma. Cancer Letters **457**, 98–109 (2019)
33. Yang, J., Lee, S.H., Goddard, M.E., Visscher, P.M.: GCTA: A Tool for Genome-wide Complex Trait Analysis. The American Journal of Human Genetics **88**(1), 76–82 (2011)