RESEARCH

BG2: Bayesian variable selection in generalized linear mixed models with nonlocal priors for non-Gaussian GWAS data

Shuangshuang Xu[1], Jacob Williams[1] and Marco A. R. Ferreira[1*]

*Correspondence:
marf@vt.edu

[1] Department of Statistics,
Virginia Tech, Blacksburg, VA
24061, USA

Abstract

Background: Genome-wide association studies (GWASes) aim to identify single nucleotide polymorphisms (SNPs) associated with a given phenotype. A common approach for the analysis of GWAS is single marker analysis (SMA) based on linear mixed models (LMMs). However, LMM-based SMA usually yields a large number of false discoveries and cannot be directly applied to non-Gaussian phenotypes such as count data.

Results: We present a novel Bayesian method to find SNPs associated with non-Gaussian phenotypes. To that end, we use generalized linear mixed models (GLMMs) and, thus, call our method Bayesian GLMMs for GWAS (BG2). To deal with the high dimensionality of GWAS analysis, we propose novel nonlocal priors specifically tailored for GLMMs. In addition, we develop related fast approximate Bayesian computations. BG2 uses a two-step procedure: first, BG2 screens for candidate SNPs; second, BG2 performs model selection that considers all screened candidate SNPs as possible regressors. A simulation study shows favorable performance of BG2 when compared to GLMM-based SMA. We illustrate the usefulness and flexibility of BG2 with three case studies on cocaine dependence (binary data), alcohol consumption (count data), and number of root-like structures in a model plant (count data).

Keywords: Bayesian statistics, GLMM, GWAS, Nonlocal prior, Variable selection

Introduction

Genome-wide association studies (GWAS) have uncovered many single nucleotide polymorphisms (SNP) associated to important phenotypes such as plant productivity [1], plant response to salt stress [2], and human diseases [3]. To take into account the correlation among GWAS observations, the most widely used methods for the analysis of GWAS continuous Gaussian data are single marker analysis (SMA) methods based on linear mixed models (LMMs) [4–6]. Recently, SMA based on logistic regression with random effects has been proposed for the analysis of GWAS binary data [7]. However, to

Xu *et al. BMC Bioinformatics*   (2023) 24:343

Page 2 of 17

the best of our knowledge, there are no published methods for the analysis of other types of *correlated GWAS non-Gaussian* data such as count data. One of our contributions is to propose the use of generalized linear mixed models for the analysis of GWAS non-Gaussian data. To that end, we use generalized linear mixed models (GLMMs) and, thus, call our method Bayesian GLMMs for GWAS (BG2).

BG2 has two steps: a screening step and a model selection step. The screening step, similarly to SMA methods, fits p GLMMs where each model has just one SNP, and uses Bayesian FDR control [8, 9] to provide a set of candidate SNPs. After that, the model selection step performs a model search through the space of GLMMs that may include any number of screened candidate SNPs as possible regressors. BG2 implements both steps using a pseudo-likelihood approach. We note that a similar pseudo-likelihood approach can be used to implement SMA methods for non-Gaussian GWAS data, and a particular case of such an approach has been proposed for GWAS binary data [7]. However, simulation studies presented in Sect. 4 show that, when compared to such SMA methods for non-Gaussian data, BG2 leads to much lower FDR.

The GLMMs for GWAS data considered by BG2 may have two types of random effects: kinship randoms effects and overdispersion random effects. The kinship random effects account for correlation among GWAS observations due to population stratification and hidden relatedness. Similarly to existing literature for Gaussian GWAS data, we assume that the vector of kinship random effects follows a multivariate Gaussian distribution with a mean vector of zeros and a covariance matrix that is the product of a one-dimensional unknown variance parameter and a known positive semi-definite kinship matrix [10, 11]. The overdispersion random effects allow for extra variability not accounted for by the model for observations; for example, when assuming a conditional Poisson model for the observations, the overdispersion random effects account for extra-Poisson variability.

Both screening and model selection steps in BG2 are based on nonlocal priors. To the best of our knowledge, this is the first time that nonlocal priors are proposed for regression coefficients in GLMMs. Previous literature in Bayesian model selection for GLMMs has assigned for regression coefficients local priors [12]. While local priors have positive density at null parameter values, nonlocal priors have density equal to zero at null parameter values. Nonlocal priors were first proposed by [13, 14] for Gaussian linear models. Nonlocal priors have been successfully developed for many different problems such as model selection in Gaussian directed acyclic graphical models [15], classification with Bayesian probit models [16], variable selection in logistic models [17], Bayesian wavelet analysis [18], and variable selection in generalized linear models [19]. In particular, [20, 21] have proposed methods based on nonlocal priors for variable selection in linear mixed models applied to GWAS data. However, because LMMs applied to binary or count data may lead to meaningless negative predictions and statistically inefficient estimation, LMMs should not be applied to non-Gaussian data such as count and binary data, which are the types of data considered by BG2. Nonlocal priors lead to faster accumulation of evidence in favor of a true null hypothesis [13], and have been shown to be advantageous for high-dimensional problems [14, 16, 22]. Therefore, BG2 uses nonlocal priors for SNP search in GWAS analysis.

Due to the large number of GLMMs that need to be fitted, BG2 relies on two approximations to speed up computations: a pseudo-likelihood approximation; and a Population Parameters Previously Determined (P3D) approximation that may be seen as an empirical Bayes approach. For GLMMs, the integrated likelihood function obtained by integrating out the random effects is not available in closed form. Repeated numerical integration of the random effects for each GLMM fitted for a GWAS analysis is computationally too expensive. Thus, BG2 uses a pseudo-likelihood approach [23] to facilitate integrating out the random effects. Such pseudo-likelihood approach leads to a Gaussian approximation for adjusted observations that allows analytically integrating out the random effects. In addition, to avoid the computation of matrix inverses for each SNP and, thus, to further speed up computations, we propose a P3D approximation for GLMMs. A P3D approximation was first proposed by [24] for Gaussian linear mixed models (LMMs) and a variation of this approximation is used in the celebrated and widely used method EMMAX for the analysis of GWAS Gaussian data [6]. With our P3D approach, BG2 needs to compute a spectral decomposition only once for each screening step and only once for each model selection step.

In our P3D approach, for each BG2 step (screening and model selection) we fit a baseline GLMM to obtain adjusted observations and estimates of the variance parameters. We then keep the adjusted observations and the variance parameters fixed at the values computed with the baseline GLMM when fitting all other models in that BG2 step. In our P3D approach, the baseline model is different for the screening step and for the model selection step. For the screening step, the baseline model is a GLMM without any SNPs. For the model selection step, the baseline model is a GLMM with all candidate SNPs obtained from the screening step. This choice of baseline GLMM for the model selection step is based on [25], who have suggested for GLMMs the use of adjusted observations based on the full model – the model with all the regressors – when computing BICs for all possible models. Therefore, BG2 with our P3D approximation does not need to compute a spectral decomposition for each SNP. As a result, when compared to a usual pseudo-likelihood approach to GLMMs, our P3D approximation greatly reduces the computational time and allows the analysis of non-Gaussian GWAS data within a reasonable time frame.

To be technical, in this work we use a hierarchical model and an empirical Bayes approach to estimate the hyperparameters of the prior distribution of the regression coefficients of GLMMs. We then combine this prior distribution with the data through Bayes Theorem to compute the posterior probability of the competing GLMMs. The Bayesian model selection procedure that we propose in this work is similar to that of [26], except that in our current work we are dealing with the problem of Bayesian ultra-high dimensional variable selection (p two orders of magnitude larger than n) in GLMMs applied to GWAS analysis. To the best of our knowledge, currently there are no published methods for Bayesian ultra-high dimensional variable selection in GLMMs.

The remainder of this paper is organized as follows. Section 2 describes the GLMMs that we consider for non-Gaussian GWAS data. Section 3 describes our BG2 method for the identification of causal SNPs. Section 4 presents the results of two simulation studies for binary data and for count data. Section 5 illustrates our method with applications to three case studies: human cocaine dependence, alcohol consumption, and the number of

Xu *et al. BMC Bioinformatics*   (2023) 24:343

Page 4 of 17

root-like structures in the plant *A. Thaliana*. Section 6 concludes with a discussion and future directions.

GLMMs for GWAS

Consider observations $y_1, \ldots, y_n$ that, given random effects, are conditionally independent and have a distribution from the exponential family of distributions. This flexible family of distributions includes the Bernoulli, binomial, Poisson, and gamma distributions. Thus, this family may be used to model observed GWAS phenotypes such as an indicator of disease presence/absence, number of lateral roots in plants, or survival time. Then, the density function of y_i is

$$f(y_i|\eta_i) = \exp[T(y_i)\eta_i - B(\eta_i) + C(y_i)], \tag{1}$$

for $i = 1, \ldots, n$, where $T(y_i)$ is the sufficient statistics for y_i, $B(.)$ and $C(.)$ are known functions. Further, each observation y_i has mean $\mu_i = B'(\eta_i)$ and variance $v_i = B''(\eta_i)$. Let X_s be a matrix of SNPs and $\boldsymbol{\beta}_s$ be the corresponding vector of regression coefficients. In addition, let X_c be a matrix that contains a column of ones for the intercept and other columns for control covariates (e.g., age, sex, and environmental factors) and $\boldsymbol{\beta}_c$ be the corresponding vector of regression coefficients. Thus, $\boldsymbol{\beta}_s$ and $\boldsymbol{\beta}_c$ are fixed effects. Further, let $\boldsymbol{\alpha}_1$ be a vector of random effects that accounts for kinship correlation. Specifically, $\boldsymbol{\alpha}_1$ has a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\kappa_1 \Sigma$, where κ_1 is an unknown scalar and Σ is a kinship matrix. Furthermore, let $\boldsymbol{\alpha}_2$ be a vector of overdispersion random effects following $N(\mathbf{0}, \kappa_2 I)$. Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be the vector of observed phenotypes. Then, the conditional expectation $E(\boldsymbol{y}|\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ is linked to the linear predictor $X_s\boldsymbol{\beta}_s + X_c\boldsymbol{\beta}_c + \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2$ by the link function g:

$$g(E(\boldsymbol{y}|\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)) = X_s\boldsymbol{\beta}_s + X_c\boldsymbol{\beta}_c + \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2. \tag{2}$$

The class of GLMMs given by Eqs. (1) and (2) can be expanded to deal with other cases. For example, to account for the experimental design used for data collection, we may add another random effect $\boldsymbol{\alpha}_3$ following a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\kappa_3 \Sigma_3$, where κ_3 is a unknown parameter and Σ_3 is a symmetric positive semi-definite matrix that describes the dependence structure among the observations due to the experimental design. Because of the P3D approach, BG2 can include additional random effects and still use the spectral decomposition approach to speed up computations.

BG2: Bayesian SNP selection in GLMMs for GWAS

Our method BG2 consists of two steps: screening and model selection. The BG2 screening step uses a novel Bayesian single marker analysis for non-Gaussian data and applies Bayesian false discovery rate control to yield a set of candidate SNPs. After that, the BG2 model selection step performs a search through the model space of all GLMMs that may include any number of SNPs from the set of candidate SNPs. In both steps, BG2 uses a pseudo-likelihood approach to fit models. In what follows, Sect. 3.1 presents the pseudo-likelihood approach, Sect. 3.2 introduces the BG2 screening step, and Sect. 3.3 presents the BG2 model selection step.

Pseudo-likelihood model fitting

In both the screening and the model selection steps, BG2 uses a pseudo-likelihood approach. In this subsection, we provide a summary description of the pseudo-likelihood approach. In addition, Additional file 1: Section S1 provides a detailed presentation of the pseudo-likelihood approach. This is an iterative approach that writes the model for the observations as $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is a vector of errors and $V = Var(\boldsymbol{\epsilon}) = Var(\boldsymbol{y}) = diag(v_1, \ldots, v_n)$ is a diagonal matrix. Note that for distributions in the exponential family, the variance v_i depends on the linear predictor η_i and, thus, gets updated in each iteration of the pseudo-likelihood algorithm. More details can be found in Additional file 1: Section S1. In addition, the pseudo-likelihood approach expands $\boldsymbol{\mu} = E(\boldsymbol{y}|\boldsymbol{\beta}_s, \boldsymbol{\beta}_c, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ in a first-order Taylor expansion around current estimates of $\boldsymbol{\beta}_s$, $\boldsymbol{\beta}_c$, $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, κ_1, and κ_2. The resulting equation is rearranged such that the left-hand side depends only on known quantities (observations, current estimates of parameters, regression matrices). Then, this equation is pre-multiplied by V^{-1}. Let $\hat{V}$ be the current estimate for V. The left-hand side of the resulting equation, known as the vector of adjusted observations, is $\boldsymbol{y}^{\star} = \widehat{V}^{-1}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}) + X_s\widehat{\boldsymbol{\beta}}_s + X_c\widehat{\boldsymbol{\beta}}_c + \widehat{\boldsymbol{\alpha}}_1 + \widehat{\boldsymbol{\alpha}}_2$. Equating $\boldsymbol{y}^{\star}$ to the right-hand side of the resulting equation yields

$$\boldsymbol{y}^{\star} = X_s\boldsymbol{\beta}_s + X_c\boldsymbol{\beta}_c + \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 + \widehat{V}^{-1}\boldsymbol{\epsilon}. \tag{3}$$

Then, the pseudo-likelihood approach approximates the GLMM by an LMM given by Eq. (3) with vectors of random effects $\boldsymbol{\alpha}_1 \sim N(\boldsymbol{0}, \kappa_1\Sigma)$ and $\boldsymbol{\alpha}_2 \sim N(\boldsymbol{0}, \kappa_2 I)$. Based on this LMM, new estimates are computed for $\boldsymbol{\beta}_s$, $\boldsymbol{\beta}_c$, $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, κ_1, κ_2, and V. The pseudo-likelihood algorithm then iterates until convergence of these estimates. More details about the pseudo-likelihood method are given in Additional file 1: Section S1.

BG2 screening step

The BG2 screening step uses a P3D approach based on a baseline model that assumes a linear predictor given in Eq. (2) specialized to contain no SNPs, that is, $g(E(\boldsymbol{y}|\boldsymbol{\beta}_c, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)) = X_c\boldsymbol{\beta}_c + \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2$.

Our P3D approach keeps $\boldsymbol{\beta}_c$, κ_1, κ_2, and V fixed at their pseudo-likelihood estimates when performing the Bayesian SMA in the BG2 screening step. Let us denote these estimates by $\widehat{\boldsymbol{\beta}}_c$, $\widehat{\kappa}_1$, $\widehat{\kappa}_2$, and $\widehat{V}$. In addition, our P3D approach keeps the vector of adjusted observations fixed equal to $\boldsymbol{y}^{\star}$ obtained at the last iteration of the pseudo-likelihood algorithm for the baseline model. Let $H = \widehat{\kappa}_1\Sigma + \widehat{\kappa}_2 I + \widehat{V}^{-1}$ be the estimated covariance matrix of the adjusted observations $\boldsymbol{y}^{\star}$. Consider the spectral decomposition of the matrix H given by $H = PDP^T$. The matrix H is kept fixed for all SNPs in the screening step. Thus, the spectral decomposition of H, which has a computational cost of $O(n^3)$, has to be computed only once at the beginning of the screening step.

Let $\boldsymbol{x}_s$ be the vector of covariate values for SNP s. Then, the BG2 screening step assumes for each SNP s, $s = 1, \ldots, p$, that the adjusted observations $\boldsymbol{y}^{\star}$ can be modeled by the LMM

$$\boldsymbol{y}^{\star} = X_c\widehat{\boldsymbol{\beta}}_c + \boldsymbol{x}_s\beta_s + \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 + \widehat{V}^{-1}\boldsymbol{\epsilon}. \tag{4}$$

Then, the adjusted observations $\boldsymbol{y}^\star$ have an approximate multivariate Gaussian distribution $N(X_c\widehat{\boldsymbol{\beta}}_c + \boldsymbol{x}_s\beta_s, H)$. Let $\widetilde{\boldsymbol{y}} = P^T(\boldsymbol{y}^\star - X_c\widehat{\boldsymbol{\beta}}_c)$ and $\widetilde{\boldsymbol{x}}_s = P^T\boldsymbol{x}_s$. Then, an estimator of β_s is $\widehat{\beta}_s = (\widetilde{\boldsymbol{x}}_s^T D^{-1}\widetilde{\boldsymbol{x}}_s)^{-1}\widetilde{\boldsymbol{x}}_s^T D^{-1}\widetilde{y}$. In addition, the estimator $\widehat{\beta}_s$ has approximate distribution $N(\beta_s, \sigma_s^2)$, where $\sigma_s^2 = var(\widehat{\beta}_s) = (\widetilde{\boldsymbol{x}}_s^T D^{-1}\widetilde{\boldsymbol{x}}_s)^{-1}$.

We assign for β_s a prior that is a mixture of a Dirac delta function and a nonlocal prior, that is,

$$p(\beta_s|\tau, \pi_0) = \pi_0\delta_0(\beta_s) + (1 - \pi_0)\frac{\beta_s^2}{n\tau\sigma_s^2}N(\beta_s|0, n\tau\sigma_s^2),$$

where π_0 is the probability of the null hypothesis that β_s is equal to zero and $\tau > 0$ is a scale parameter. Larger values of τ cause the prior to be more spread out and lead BG2 to focus on identifying SNPs with relatively large regression coefficients. Then, the predictive density of $\widehat{\beta}_s$ is

$$\begin{aligned}
p(\widehat{\beta}_s|\tau, \pi_0) &= \int p(\widehat{\beta}_s|\beta_s)p(\beta_s|\tau, \pi_0)\, d\beta_s \\
&= \pi_0 N(\widehat{\beta}_s|0, \sigma_s^2) + (1 - \pi_0)(2\pi\sigma_s^2)^{-\frac{1}{2}}(n\tau + 1)^{-\frac{3}{2}} \\
&\quad \exp\left\{-\frac{\widehat{\beta}_s^2}{2\sigma_s^2(n\tau + 1)}\right\}\left[1 + \frac{n\tau\widehat{\beta}_s^2}{(n\tau + 1)\sigma_s^2}\right].
\end{aligned} \tag{5}$$

Based on this predictive density and assuming that $\widehat{\beta}_1, \ldots, \widehat{\beta}_p$ are approximately conditionally independent given π_0 and τ, we obtain the approximate likelihood function of τ and π_0

$$L(\widehat{\beta}_1, \ldots, \widehat{\beta}_p|\tau, \pi_0) = \prod_{s=1}^{p} p(\widehat{\beta}_s|\tau, \pi_0). \tag{6}$$

Let $\pi(\tau)$ and $\pi(\pi_0)$ be the prior densities of τ and π_0, respectively. Then, by Bayes Theorem an approximate posterior density for (τ, π_0) is

$$\pi(\tau, \pi_0|\widehat{\beta}_1, \ldots, \widehat{\beta}_p) \propto \pi(\tau)\pi(\pi_0)\prod_{s=1}^{p} p(\widehat{\beta}_s|\tau, \pi_0). \tag{7}$$

BG2 estimates τ and π_0 by maximizing (7) to obtain posterior modes $\widehat{\tau}$ and $\widehat{\pi}_0$.

We assign a noninformative uniform prior on (0, 1) for π_0 and consider two prior specifications for τ. The first prior specification is a uniform prior for τ on $(0, \infty)$. The second prior specification for τ is an inverse gamma distribution with shape parameter $0.55/0.022 + 1$ and rate parameter 0.55, that is $\tau \sim IG(0.55/0.022 + 1, 0.55)$. This prior specification implies a prior mean for τ equal to 0.022, which was the value for a fixed τ recommended by [20] for GWAS studies. In addition, we note that values of τ that are too small lead to numerical instability. Therefore, our prior specification implies that *a priori* $P(\tau > 0.01) = 0.999$, stochastically keeping τ away from 0.

Alternatively, we may fix τ at pre-specified values [14, 20]. Specifically, in the context of GWAS analysis, [20] suggested fixing $\tau = 0.022$ because GWAS effect sizes are generally very small. When $\tau = 0.022$, the nonlocal product moment prior (pMOM) prior assigns a probability of 0.01 to the event that a standardized effect size falls in the interval $(-$

0.05, 0.05). Thus, in the simulation studies presented in Sect. 4, we also consider fixing τ at 0.022.

After estimating τ and π_0, BG2 takes an Empirical Bayes approach and keep them at their estimates $\widehat{\tau}$ and $\widehat{\pi}_0$ while using Bayes Theorem to compute the posterior probability that the regression coefficient of SNP s $(s = 1, \ldots, p)$ in the screening step is different than zero, that is

$$P(\beta_s \neq 0 | \widehat{\beta}_s, \widehat{\tau}, \widehat{\pi}_0) = 1 - \frac{\widehat{\pi}_0 N(\widehat{\beta}_s | 0, \sigma_s^2)}{p(\widehat{\beta}_s | \widehat{\tau}, \widehat{\pi}_0)}, \tag{8}$$

where $p(\widehat{\beta}_s | \widehat{\tau}, \widehat{\pi}_0)$ is the predictive density given in Eq. (5) with $\tau = \widehat{\tau}$ and $\pi_0 = \widehat{\pi}_0$.

Finally, based on the posterior probabilities computed with Eq. (8), the BG2 screening step uses Bayesian FDR control [8, 9, 27–29] to obtain a list of candidate SNPs while keeping the nominal FDR at 5%. Let us denote the number of SNPs contained in this list of candidate SNPs obtained in the screening step by k.

BG2 model selection step

The BG2 model selection step considers GLMMs with any number of SNPs from the list of k candidate SNPs obtained from the BG2 screening step. Thus, the model selection step considers $S = 2^k$ possible models. Let M_m be the m-th model, $m = 1, \ldots, S$. Let X_m be the matrix of SNPs in model M_m, $\boldsymbol{\beta}_m$ be the corresponding vector of regression coefficients, and p_m be the number of SNPs in model M_m. Let X_S be the model with all k candidate SNPs.

We assume that the k candidate SNPs may or may not be in a model according to a sequence of exchangeable Bernoulli trials. Specifically, the prior probability of model M_m is $P(M_m) = \widehat{\pi}_0^{k-p_m}(1 - \widehat{\pi}_0)^{p_m}$ where $\widehat{\pi}_0$ is the estimate of the probability of null hypothesis obtained in the screening step. We do this to ensure that the Bayesian control of false discoveries in the BG2 model selection step is as strict as the control of false discoveries in the BG2 screening step.

The BG2 model selection step uses a P3D approach where the baseline model is the full model M_S with linear predictor $g(E(\boldsymbol{y}|\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)) = X_c\boldsymbol{\beta}_c + X_S\boldsymbol{\beta}_S + \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2$. The pseudo-likelihood approach then yields estimates $\widehat{\boldsymbol{\beta}}_c$, $\widehat{\kappa}_1$ and $\widehat{\kappa}_2$, $\widehat{V}$, and adjusted observations $\boldsymbol{y}^\star$. We then consider all models $M_m, m = 1, \ldots, S$, where we keep $\boldsymbol{\beta}_c$, κ_1, κ_2, and V fixed at these estimates. Let $H = \widehat{\kappa}_1 \Sigma + \widehat{\kappa}_2 I + \widehat{V}^{-1}$ and consider the spectral decomposition of the matrix H given by $H = PDP^T$. The matrix H is kept fixed for all non-baseline models in the model selection step. Thus, even though the spectral decomposition has a computational cost of $O(n^3)$, this decomposition has to be computed only once at the beginning of the model selection step. In addition, following the recommendation of [25], we keep the adjusted observations for all the S considered models fixed at the adjusted observations $\boldsymbol{y}^\star$ obtained while fitting the full model.

Therefore, under model M_m and with the P3D approach, the adjusted observations $\boldsymbol{y}^\star$ follow the approximate distribution $N\left(X_c\widehat{\boldsymbol{\beta}}_c + X_m\boldsymbol{\beta}_m, H\right)$. In addition, let $\widetilde{\boldsymbol{y}} = P^T(\boldsymbol{y}^\star - X_c\widehat{\boldsymbol{\beta}}_c)$ and $\widetilde{X}_m = P^T X_m$. Then, we can rewrite the LMM as $\widetilde{\boldsymbol{y}}|\boldsymbol{\beta}_m \overset{a}{\sim} N(\widetilde{X}_m\boldsymbol{\beta}_m, D)$, where $\overset{a}{\sim}$ denotes "approximately distributed as." Because D is a diagonal matrix, computations for this latter model are very fast.

We propose a novel nonlocal prior for GLMMs. Specifically, we propose a prior density that is the product of a multivariate Gaussian density and the product of the square of each element of the vector of regression coefficients β_m. In this multivariate Gaussian density, the covariance matrix is $\tau n(X_m^T H^{-1} X_m)^{-1}$. Using the spectral decomposition of the matrix H, the prior we propose for $\boldsymbol{\beta}_m$ is

$$
\pi(\boldsymbol{\beta}_m|M_m) = d_m(2\pi)^{-p_m/2}(\widehat{\tau}n)^{-3p_m/2}|\widetilde{X}_m^T D^{-1}\widetilde{X}_m|^{\frac{3}{2}}
$$
$$
\exp\left[-\frac{1}{2\widehat{\tau}n}\boldsymbol{\beta}_m^T \widetilde{X}_m^T D^{-1}\widetilde{X}_m\boldsymbol{\beta}_m\right]\prod_{i=1}^{p_m}\beta_{mi}^2, \tag{9}
$$

where d_m is a normalizing constant.

Let $C_m = \widetilde{X}_m^T D^{-1}\widetilde{X}_m(1 + (\widehat{\tau}n)^{-1})$, $\widetilde{\boldsymbol{\beta}}_m = C_m^{-1}\widetilde{X}_m^T D^{-1}\widetilde{\boldsymbol{y}}$, and $R_m = \widetilde{\boldsymbol{y}}^T D^{-1}(D - \widetilde{X}_m C_m^{-1}\widetilde{X}_m^T)$ $D^{-1}\widetilde{\boldsymbol{y}} = \widetilde{\boldsymbol{y}}^T D^{-1}\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{y}}^T D^{-1}\widetilde{X}_m\widetilde{\boldsymbol{\beta}}_m$. Then, the marginal density of the adjusted observations $\widetilde{\boldsymbol{y}}$ conditional on model M_m is

$$
m(\widetilde{\boldsymbol{y}}|M_m) = \int N(\widetilde{\boldsymbol{y}}|\widetilde{X}_m\boldsymbol{\beta}_m, D)\pi(\boldsymbol{\beta}_m|M_m)\,d\boldsymbol{\beta}_m
$$
$$
= (2\pi)^{-\frac{n}{2}}|D|^{-\frac{1}{2}}(1 + \widehat{\tau}n)^{-p_m/2} \tag{10}
$$
$$
\exp\left(-\frac{R_m}{2}\right)\frac{E_2\left(\prod_{i=1}^{p_m}\beta_{mi}^2\right)}{E_1\left(\prod_{i=1}^{p_m}\beta_{mi}^2\right)},
$$

where $E_1\left(\prod_{i=1}^{p_m}\beta_{mi}^2\right)$ is the expected value with respect to $N(\boldsymbol{0}, (1 + \widehat{\tau}n)C_m^{-1})$ and $E_2\left(\prod_{i=1}^{p_m}\beta_{mi}^2\right)$ is the expected value with respect to $N(\widetilde{\boldsymbol{\beta}}_m, C_m^{-1})$. To approximate $E_1\left(\prod_{i=1}^{p_m}\beta_{mi}^2\right)$ and $E_2\left(\prod_{i=1}^{p_m}\beta_{mi}^2\right)$, we simulate 2000 samples from $N(\widetilde{\boldsymbol{\beta}}_m, C_m^{-1})$, denoted as $\boldsymbol{\beta}_{2m}^{(j)}$, $j = 1, \ldots, 2000$. We compute $\sum_{j=1}^{2000}(\prod_{i=1}^{p_m}\beta_{2mi}^{2(j)})/2000$ as an approximation to $E_2\left(\prod_{i=1}^{p_m}\beta_{mi}^2\right)$. Let $\boldsymbol{\beta}_{1m}^{(j)} = (1 + \widehat{\tau}n)^{\frac{1}{2}}(\boldsymbol{\beta}_{2m}^{(j)} - \widetilde{\boldsymbol{\beta}}_m)$, $j = 1, \ldots, 2000$. Finally, we compute $\sum_{j=1}^{2000}(\prod_{i=1}^{p_m}\beta_{1mi}^{2(j)})/2000$ as an approximation to $E_1\left(\prod_{i=1}^{p_m}\beta_{mi}^2\right)$.

Then, the posterior probability of model M_m is

$$
P(M_m|\widetilde{\boldsymbol{y}}) \propto P(M_m)m(\widetilde{\boldsymbol{y}}|M_m). \tag{11}
$$

Note that the posterior distribution of the vector of regression coefficients is multimodal. BG2 deals with this multimodality without any difficulties. In the screening step, when β_s is a scalar, we compute the posterior probability of $\beta_s \neq 0$ using Eqs. (5) and (8). In the model selection step, when $\boldsymbol{\beta}_m$ is a vector of coefficients, we compute the posterior probability of model M_m using Eqs. (10) and (11).

If the number of candidate covariates k is small ($k < 16$), we compute the posterior probabilities for all 2^k candidate models and select the highest posterior probability model as the best model. If the number of candidate covariates is large, we use a genetic algorithm from the R package GA [30] to search for the highest posterior probability model.

Simulation studies

We have performed simulation studies to compare our nonlocal-prior-based BG2 method versus SMA for binary data and count data. Specifically, we consider single marker analysis with Bonferroni correction with nominal FDR set to 0.05. To assess the

performance of our methods, in these simulation studies we use genotype SNP data from humans and from A. Thaliana. These are the same genotype data used in the case studies we present in Sect. 5. We use four criteria to compare the competing methods: true positives (TP), false positives (FP), false discovery rate (FDR) and F1 score. Within each simulation study, for each method we compute the average TP, FP, FDR and F1 over 100 simulated datasets. We use a buffer to define what is a true positive and a false positive. Following [21], if one or more detected SNPs are adjacent (within 5000 base pairs) to a same causal SNP, that is counted as a true positive. In addition, each detected SNP not adjacent to a causal SNP is counted as a false positive.

Binary data

We simulate binary GWAS data using genotype information from the Study of Addiction: Genetics and Environment (SAGE) which is part of the National Human Genome Research Institute's Gene Environment Association Study Initiative [Database for Genotypes and Phenotypes (dbGaP) study accession phs000092.v1.p1]. Specifically, we use genotype information from 2,772 European Americans in a total of 800,000 SNPs with minor allele frequency (MAF) larger than 0.01.

From these 800,000 SNPs, we selected 20 evenly spaced SNPs to be the causal SNPs. We set the regression coefficients for 5 of these causal SNPs to 0.2, and for 5 other causal SNPs to −0.2. In addition, the regression coefficients for the other 10 causal SNPs have the same value β, but that value varies in six settings: 0.2, 0.3, 0.4, 0.5, 0.7 and 1. Further, we set the intercept at $\beta_0 = -0.5$. Furthermore, the variance component κ of the kinship random effects $\boldsymbol{\alpha}$ is set to 0.15. Thus, the binary phenotype data are simulated from a Bernoulli GLMM with logistic link function and linear predictor $\beta_0 + \sum_{i=1}^{10} \beta x_{ij} + \sum_{i=11}^{15} 0.2 x_{ij} + \sum_{i=16}^{20} (-0.2) x_{ij} + \alpha_i$, with $\boldsymbol{\alpha} \sim N(\mathbf{0}, \kappa \Sigma)$ where Σ is the kinship matrix.

Figure 1 shows for binary data the performance of our BG2 method with three different ways to choose the parameter τ, as well as the performance of the SMA method. These performances in terms of TP, FP, FDR, and F1 averaged over 100 datasets for each setting are plotted as functions of the varying regression coefficient β. In addition, Fig. 1 shows the computational time. Additional file 1: Figure S2 show boxplots of TP, FP, FDR, and F1. Our BG2 methods take twice as long as SMA, which is to be expected since SMA has only a screening step whereas BG2 has a screening step and a model selection step. Among the three ways considered to choose τ for BG2, estimating τ based on a uniform prior provides higher F1 scores for smaller values of β, and provides comparable F1 scores for larger values of β. In addition, when compared to SMA, BG2 with uniform prior provides larger average number of true positives TP than when β is small, and a smaller TP when β is large. However, BG2 with uniform prior leads to a much smaller average number of false positives than SMA. As a result, when compared to SMA, for all considered values of the regression coefficient β, BG2 with uniform prior has much larger F1.

Finally, we have tested the robustness of BG2 to the case of binary GWAS data with no causal SNPs. Specifically, we have simulated 100 datasets with binary GWAS data from a Bernoulli GLMM with logit link function and linear predictor $\beta_0 + \alpha_i$. While BG2 with any of the ways to choose τ does not yield any false positive for 100 simulated datasets,

(a) True positives

(b) False positives

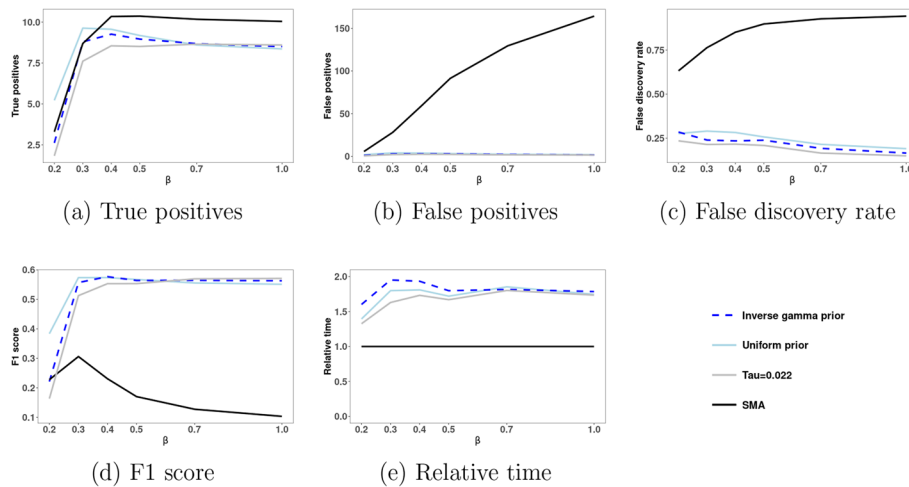(c) False discovery rate

(d) F1 score

(e) Relative time

Fig. 1 SNP search performance of BG2 and SMA methods for simulated binary data

SMA has an average of 0.06 false positives. Therefore, BG2 performs better than SMA for binary GWAS data and is robust to the case when there are no causal SNPs.

Count data

We simulate count GWAS data using genotype information from The Arabidopsis Information Resource (TAIR9) (https://www.arabidopsis.org/). This simulation study is based on a case study on root-like structures in *A. Thaliana* that we present in Sect. 5.3.

Specifically, we use 188,980 SNPs with MAF>0.01 from 152 ecotypes of *A. Thaliana*. This simulation study assumes 10 causal SNPs evenly located among all available SNPs. Of these 10 causal SNPs, 5 causal SNPs have fixed coefficients equal to 0.2, and the other 5 causal SNPs have the same coefficient β which varies in eight settings: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7 and 1. In addition, we set the intercept β_0 equal to 1. Further, we assume that there are two random effects: a kinship random effect $\boldsymbol{\alpha}_1$ with variance component κ_1 equal to 1; and an overdispersion random effect $\boldsymbol{\alpha}_2$ with variance component κ_2 equal to 0.3, which is close to the estimate obtained in the case study presented in Sect. 5.3. Let r_i be the number of replicates of ecotype i. Because in the case study most ecotypes have 12 replicates, in this simulation study we assume that all ecotypes have 12 replicates. In addition, the phenotype y_i for ecotype i is the total number of root-like structures of the r_i replicates. These phenotype count data are sampled from a Poisson GLMM with logarithm link function and linear predictor $\log(r_i) + \beta_0 + \sum_{i=1}^{5} \beta x_{ij} + \sum_{i=6}^{10} 0.2x_{ij} + \alpha_{1i} + \alpha_{2i}$.

Figure 2 shows for count data the performance of our BG2 method as well as the performance of the SMA method. These performances are averaged over 100 simulated datasets for each setting and plotted as functions of the varying regression coefficient β. In addition, Fig. 2 shows the computational time. Additional file 1: Figure S3 show boxplots of TP, FP, FDR, and F1. Our BG2 methods take about eight times longer than SMA, but they still provide results in a feasible amount of time. Among the three ways considered to choose τ for BG2, estimating τ based on an inverse gamma prior provides larger average number of true positives and about the same FDR level. As a result, when
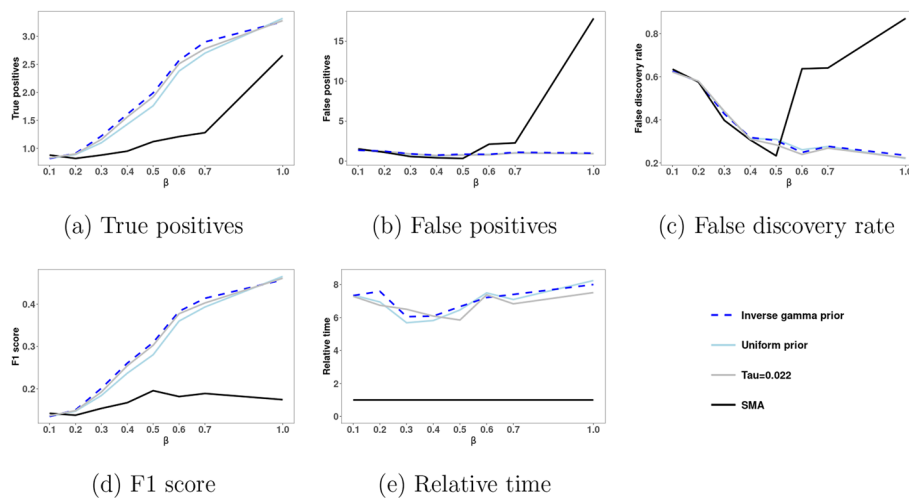
(a) True positives   (b) False positives   (c) False discovery rate

(d) F1 score   (e) Relative time

Fig. 2 SNP search performance of BG2 and SMA methods for simulated count data

compared to the other ways to choose τ, estimating τ based on an inverse gamma prior has higher F1 scores for most considered values of β. In addition, when compared to SMA, BG2 with an inverse gamma prior provides larger average number of true positives TP for most considered values of β. Further, BG2 with inverse gamma prior has about the same FDR level as SMA for $\beta \leq 0.5$ and a much smaller FDR level for $\beta > 0.5$. As a result, while BG2 with an inverse gamma prior has comparable F1 to SMA for small values of β, the F1 of BG2 with an inverse gamma prior becomes much larger than the F1 of SMA as β increases.

In addition, we have tested the robustness of BG2 to the case of count GWAS data with no causal SNPs. Specifically, we have simulated 100 datasets with count GWAS data from a Poisson GLMM with logarithm link function and linear predictor $\beta_0 + \alpha_{1i} + \alpha_{2i}$. The average number of false positives for all considered methods is 0. Thus, both SMA and BG2 methods perform well in the case of count GWAS data with no causal SNPs.

Choice of prior for τ

Choice of priors is an important part of the implementation of Bayesian methods. To obtain more information about the impact of prior choice on the results of BG2 implementations, we have expanded our simulation studies presented in Sects. 4.1 and 4.2. Specifically, we have performed two additional simulation studies: one that uses count data simulated with human genome and another one that uses binary data simulated with A. Thaliana genome. Unfortunately, the simulated binary datasets simulated with A. Thaliana genome did not contain enough information for SNPs to be detected by SMA or BG2. Additional file 1: Figure S1 presents the results for count data based on human genome data. Similarly to the results from Sects. 4.1 and 4.2, any of the implementations of BG2 perform much better than SMA in terms of FDR and F1. In addition, for count data based on human genome data, the three implementations of BG2 provide similar results. Therefore, from the results of the simulation studies, there is no prior choice that dominates the other prior choices.

Another important consideration is that the performance of the priors will vary for different GWAS datasets. And, of course, BG2 is to be used by scientists who are not Bayesian statisticians – thus, ideally there should be a default prior that would be safe to use with any GWAS dataset. One such class of priors is the class of non-informative priors [31–34] that impart little or no information in the analysis. The uniform prior is not concentrated around any particular value of τ and, in this sense, is non-informative. In addition, in the simulation studies considered here, BG2 with the uniform prior for τ performed similarly or better than BG2 with other prior choices for τ. Therefore, in the implementation of BG2 we recommend the uniform prior as a default choice for τ.

Case studies

To illustrate the usefulness and flexibility of the nonlocal-prior-based BG2 method, this section presents three case studies on cocaine dependence, alcohol consumption, and number of root-like structures in *A. Thaliana*.

Maximum number of alcoholic drinks

The Collaborative Study on the Genetics of Alcoholism (COGA) [35] was a large-scale family study that had as primary objective to identify genes related to alcohol dependence. Here, we consider as the response variable the maximum number of alcoholic drinks consumed in 24 h. We analyze data on 2759 European Americans considering 846,076 SNPs with MAF>0.01 and with less than 5% missing. To perform this analysis, we use the Poisson GLMM for count data considered in Sect. 4.2. In our analysis, the 846,076 SNPs are possible regressors. Our Poisson GLMM accounts for genetic structure among 2759 subjects by including a vector of kinship random effects, and allows for extra-Poisson variability with a vector of overdispersion random effects.

While SMA detected 10 SNPs, BG2 detected only one SNP. More specifically, the screening step of BG2 identified 10 candidate SNPs which were then given to the BG2 model selection step. The BG2 model selection step then identified one SNP. Likelihood ratio tests indicate that the identified SNPs do not violate the hypothesis of Hardy-Weinberg equilibrium. While we cannot be sure about which of these SNPs are false positives, the simulation studies in Sect. 4 show that SMA tends to have a much higher FDR than BG2. Thus, in this case study the nonlocal-prior-based BG2 method provides a list of SNPs for further investigation that is much more focused. The SNP detected by BG2 is located in the protein-coding gene PTGER4 on chromosome 5. The protein encoded by PTGER4 is a receptor for prostaglandin E2 (PGE2). An increase in PGE2 is part of the inflammatory response to alcohol consumption, and the use of the PGE2-inhibitor tolfenamic acid significantly reduces the severity of several hangover symptoms [36].

Cocaine dependence

In this case study, we analyze the association between cocaine dependence and single nucleotide polymorphisms (SNPs). We analyze data from the Family Study of Cocaine Dependence (FSCD) [37], which was part of the Study of Addiction: Genetics and Environment. Specifically, we analyze data on 2,767 European Americans considering 846,076 SNPs with MAF>0.01 and with less than 5% missing. The response variable is whether or not the subject is addicted to cocaine. To perform this analysis, we use the

model for binary data considered in Sect. 4.1. Because males and females seem to have different behaviors with respect to cocaine use, we include sex as a control covariate. All 846,076 SNPs are possible regressors. In addition, to account for the genetic structure among the 2767 subjects, our Bernoulli GLMM has a vector of kinship random effects.

BG2 detects one SNP, which is located in the protein-coding gene ABCC8 on chromosome 11. In this case study, the screening step of BG2 identified 1 candidate SNP which was then selected in the model selection step. For this dataset, SMA only detects the same SNP. A likelihood ratio test indicates that the identified SNP does not violate the hypothesis of Hardy-Weinberg equilibrium. The protein encoded by this gene is a member of the superfamily of ATP-binding cassette (ABC) proteins which transport various molecules across extra-cellular and intra-cellular membranes. In addition, a quantitative transcriptomics analysis (RNA-Seq) has shown that this gene is overexpressed in the brain [38]. Further, cocaine increases expression of ABCC1 (another gene that encodes an ABC protein) in mice [39]. Finally, ABCC1-siRNA (a silencer of ABCC1) blocks cocaine-induced place preference in mice [39].

Root-like structures in *A. Thaliana*

To illustrate the application of our method to count data, we analyze data from a study of plant regeneration from root explants of the selfing species *A. Thaliana* [40]. Specifically, we consider as response variable the number of root-like structures. We note that [40] applied a square root transformation to analyze this count phenotype variable. In contrast, we use the Poisson GLMM with overdispersion considered in Sect. 4.2 to analyze the original count data. Our model contains a vector of kinship random effects to account for the correlation among individuals and a vector of overdispersion random effects. We focus on the number of root-like structures after 21 days in which seedlings are under warm white light at 21^oC following a 14/10 h light/dark regime. There are 188,980 SNPs for 152 ecotypes, with 12 biological replicates per ecotype, from TAIR9 with MAF>0.01.

BG2 detects 3 SNPs. More specifically, the screening step of BG2 screened 5 candidate SNPs and then the BG2 model selection step identified 3 of these SNPs. For this dataset, SMA detects the same 3 SNPs. These 3 SNPs are expressed in the root and are located in protein-coding genes AT1G20090, AT1G20100 and AT1G20720. Specifically, AT1G20100 encodes a DNA ligase-like protein involved in the regulation of metabolic processes. In addition, gene AT1G20720 encodes a RAD3-like DNA binding helicase protein that acts in the repair of double-strand breaks in DNA, and in nucleotide-excision repair. Finally, AT1G20090 encodes a ROP2 protein which is known to effect root hair initiation and tip growth [41].

Discussion

We have proposed BG2, a two-stage Bayesian SNP detection method for non-Gaussian GWAS data. BG2 uses a GLMM framework that includes kinship random effects and overdispersion random effects. BG2 has two steps: a screening step and a model selection step. The screening step performs a Bayesian SMA that selects a set of candidate SNPs. The model selection step then considers all possible GLMMs based on this set of candidate SNPs. To speed up computations, we develop a pseudo likelihood approach

combined with P3D. Further, we develop a novel class of nonlocal priors for the regression coefficients specially tailored for GLMMs. Simulation studies show that, for both binary and count GWAS data, BG2 is much better than SMA in terms of FDR and F1.

The simulation studies show that, when compared to SMA, BG2 has a much lower FDR. Of course, there are some combinations of parameters for which SMA and BG2 provide similar results, and that is what seems to have happened in Sects. 5.2 and 5.3. However, in some applications BG2 provides a much smaller number of false discoveries than SMA, and that is what seems to have happened in the case study presented in Sect. 5.1. Therefore, when compared to SMA, BG2 is more robust and precise.

A relevant question is how sensitive to the choice of prior is the performance of BG2. We considered here implementations of BG2 with three different choices of priors. The simulation studies presented in this paper and in the additional file have shown that BG2 implementations with each of these three different choices of priors have similar performance. As a matter of fact, BG2 with the different choices of priors considered here provide the same results for the three case studies. Therefore, the performance of BG2 seems to be relatively robust to the choice of priors.

While we have chosen to implement BG2 with a pseudo-likelihood approach and a genetic algorithm to explore the model space for the analysis of non-Gaussian GWAS data, we acknowledge that other approaches may be possible. For example, instead of the pseudo-likelihood approach, researchers may consider variational inference approaches [42, 43]. In addition, instead of combining the pseudo-likelihood approach and a genetic algorithm, researchers may consider implementing a parallel tempering approach [44] to perform estimation and model selection at the same time. However, we think that such parallel tempering approach may not be computationally feasible for Bayesian ultra-high dimensional variable selection in GLMMs applied to non-Gaussian GWAS.

There are several possible avenues for future research. One promising research direction is to adapt BG2 for application to biobank scale data. Another possible research direction is to implement BG2 with an iterative procedure that would allow smaller effect sizes to be detected. Finally, another possible research avenue is to develop BG2 for GWAS analysis when the phenotype is survival time.

Conclusion

We propose BG2, a novel two-stage Bayesian approach for non-Gaussian GWAS data. Compared to SMA, BG2 provides a much lower FDR, is more precise and robust. BG2 is implemented in the R package BG2 that is available on BioConductor at https://bioconductor.org/packages/release/bioc/html/BG2.html.

Abbreviations
BG2   Bayesian GLMMs for GWAS
LMM   Linear mixed model
GLMM   Generalized linear mixed model
SMA   Single marker analysis
GWAS   Genome-wide association studies
SNP   Single nucleotide polymorphism
TP   True positive
FP   False positive
FDR   False discovery rate
GA   Genetic algorithm
MAF   Minor allele frequency

P3D   Population parameters previously determined
MCMC   Markov chain Monte Carlo

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-023-05468-w.

> **Additional file 1**. Supplementary Material for BG2: Bayesian variable selection in generalized linear mixed models with nonlocal priors for non-Gaussian GWAS data.

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

Availability of data and materials
BG2 is implemented in the R package BG2 that is available on BioConductor at https://bioconductor.org/packages/release/bioc/html/BG2.html. In addition, the case study with A. Thaliana can be reproduced using the information and R code available at https://marf-at-vt.github.io/BG2-CaseStudy.html. The A. Thaliana phenotype data and genotype data are available from the following sources: A. Thaliana phenotype data available at https://arapheno.1001genomes.org; A. Thaliana genotype dataset available from R package qtcat.data (https://rdrr.io/github/QTCAT/qtcat.data/). Genotype and phenotype data for alcohol and cocaine use disorder in humans is available from the NIH dbGap website: https://www.ncbi.nlm.nih.gov/gap/, the accession number is phs000092.v1.p1.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

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

References
1. Zhang X, Ding W, Xue D, Li X, Zhou Y, Shen J, Feng J, Guo N, Qiu L, Xing H, et al. Genome-wide association studies of plant architecture-related traits and 100-seed weight in soybean landraces. BMC Genomic Data. 2021;22(1):1–14.
2. Julkowska MM, Koevoets IT, Mol S, Hoefsloot H, Feron R, Tester MA, Keurentjes JJ, Korte A, Haring MA, de Boer G-J, et al. Genetic components of root architecture remodeling in response to salt stress. Plant Cell. 2017;29(12):3198–213.

3.   Zhang H, Ahearn TU, Lecarpentier J, Barnes D, Beesley J, Qi G, Jiang X, O'Mara TA, Zhao N, Bolla MK, et al. Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. Nat Genet. 2020;52(6):572–81.

4.   Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 2006;38(2):203–8.

5.   Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. Efficient control of population structure in model organism association mapping. Genetics. 2008;178(3):1709–23.

6.   Kang H.M, Sul J.H, Service S.K, Zaitlen N.A, Kong S.-Y, Freimer N.B, Sabatti C, Eskin E. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010;42(4):348–54.

7.   Jiang L, Zheng Z, Fang H, Yang J. A generalized linear mixed model association tool for biobank-scale data. Nat Genet. 2021;53(11):1616–21.

8.   Newton MA, Noueiry A, Sarkar D, Ahlquist P. Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics. 2004;5(2):155–76.

9.   Muller P, Parmigiani G, Rice K. FDR and Bayesian multiple comparisons rules. In: Bernardo, JM, Bayarri MJ, Berger JO, Dawid AP, Heckerman D, Smith AFM, West M (eds) Bayesian Statistics 8, pp. 349–370. Oxford Univ. Press, Oxford; 2007.

10. Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y, Dreisigacker S, Crossa J, Sánchez-Villeda H, Sorrells M, et al. Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Genome. 2012;5(3).

11. Müller D, Technow F, Melchinger AE. Shrinkage estimation of the genomic relationship matrix can improve genomic estimated breeding values in the training set. Theor Appl Genet. 2015;128(4):693–703.

12. Baragatti M, Pommeret D. A study of variable selection using g-prior distribution with ridge parameter. Comput Stat Data Anal. 2012;56(6):1920–34.

13. Johnson VE, Rossell D. On the use of non-local prior densities in Bayesian hypothesis tests. J R Stat Soc: Ser B. 2010;72(2):143–70.

14. Johnson VE, Rossell D. Bayesian model selection in high-dimensional settings. J Am Stat Assoc. 2012;107(498):649–60.

15. Altomare D, Consonni G, La Rocca L. Objective Bayesian search of Gaussian directed acyclic graphical models for ordered variables with non-local priors. Biometrics. 2013;69(2):478–87.

16. Rossell D, Telesca D, Johnson VE. High-dimensional Bayesian classifiers using non-local priors. In: Giudici P, Ingrassia S, Vichi M, editors. Statistical models for data analysis. London: Springer; 2013. p. 305–13.

17. Nikooienejad A, Wang W, Johnson VE. Bayesian variable selection for binary outcomes in high-dimensional genomic studies using non-local priors. Bioinformatics. 2016;32(9):1338–45.

18. Sanyal N, Ferreira MAR. Bayesian wavelet analysis using nonlocal priors with an application to FMRI analysis. Sankhya B. 2017;79(2):361–88.

19. Wu H-H, Ferreira MAR, Elkhouly M, Ji T. Hyper nonlocal priors for variable selection in generalized linear models. Sankhya A. 2020;82(1):147–85.

20. Sanyal N, Lo M-T, Kauppi K, Djurovic S, Andreassen OA, Johnson VE, Chen C-H. GWASinlps: non-local prior based iterative SNP selection tool for genome-wide association studies. Bioinformatics. 2019;35(1):1–11.

21. Williams J, Xu S, Ferreira MAR. BGWAS: Bayesian variable selection in linear mixed models with nonlocal priors for genome-wide association studies. BMC Bioinform. 2023;24:194.

22. Rossell D, Telesca D. Nonlocal priors for high-dimensional estimation. J Am Stat Assoc. 2017;112(517):254–65.

23. Wolfinger R, O'Connell M. Generalized linear mixed models: a pseudo-likelihood approach. J Stat Comput Simul. 1993;48(3–4):233–43.

24. Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, et al. Mixed linear model approach adapted for genome-wide association studies. Nat Genet. 2010;42(4):355–60.

25. Ten Eyck P, Cavanaugh JE. An alternate approach to pseudo-likelihood model selection in the generalized linear mixed modeling framework. Sankhya B. 2018;80(1):98–122.

26. Xu S, Ferreira MAR, Porter EM, Franck C. Bayesian model selection for generalized linear mixed models. Biometrics. 2023.

27. Cui S, Guha S, Ferreira MAR, Tegge AN. hmmseq: a hidden Markov model for detecting differentially expressed genes from RNA-seq data. Ann Appl Stat. 2015;9(2):901–25.

28. Xie J, Ji T, Ferreira MAR, Li Y, Patel BN, Rivera RM. Modeling allele-specific expression at the gene and SNP levels simultaneously by a Bayesian logistic mixed regression model. BMC Bioinform. 2019;20(1):1–13.

29. Williams J, Ferreira MAR, Ji T. BICOSS: Bayesian iterative conditional stochastic search for GWAS. BMC Bioinform. 2022;23:475.

30. Scrucca L. GA: a package for genetic algorithms in R. J Stat Softw. 2013;53(4):1–37.

31. Berger JO, De Oliveira V, Sansó B. Objective Bayesian analysis of spatially correlated data. J Am Stat Assoc. 2001;96(456):1361–74.

32. Fonseca TCO, Ferreira MAR, Migon HS. Objective Bayesian analysis for the student-*t* regression model. Biometrika. 2008;95(2):325–33.

33. Ferreira MAR, Salazar E. Bayesian reference analysis for exponential power regression models. J Stat Distrib Appl. 2014;1(1):1–12.

34. Keefe MJ, Ferreira MAR, Franck CT. Objective Bayesian analysis for Gaussian hierarchical models with intrinsic conditional autoregressive priors. Bayesian Anal. 2019;14:181–209.

35. Begleiter H, Reich T, Hesselbrock V, Porjesz B, Li T-K, Schuckit MA, Edenberg HJ, Rice JP, et al. The collaborative study on the genetics of alcoholism. Alcohol Health Res World. 1995;19:228–228.

36. van de Loo AJ, Mackus M, Kwon O, Krishnakumar IM, Garssen J, Kraneveld AD, Scholey A, Verster JC. The inflammatory response to alcohol consumption and its role in the pathology of alcohol hangover. J Clin Med. 2020;9(7):2081.

37. Bierut LJ, Strickland JR, Thompson JR, Afful SE, Cottler LB. Drug use and dependence in cocaine dependent subjects, community-based individuals, and their siblings. Drug Alcohol Depend. 2008;95(1–2):14–22.

38. Fagerberg L, Hallström BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, Habuka M, Tahmasebpoor S, Danielsson A, Edlund K, et al. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. Mol Cell Proteomics. 2014;13(2):397–406.
39. Chen L, Chen H, Xing Y, Li J. ABCC1 regulates cocaine-associated memory, spine plasticity and GluA1 and GluA2 surface expression. NeuroReport. 2021;32(10):833–9.
40. Lardon R, Wijnker E, Keurentjes J, Geelen D. The genetic framework of shoot regeneration in Arabidopsis comprises master regulators and conditional fine-tuning factors. Commun Biol. 2020;3(1):1–13.
41. Jones MA, Shen J-J, Fu Y, Li H, Yang Z, Grierson CS. The Arabidopsis Rop2 GTPase is a positive regulator of both root hair initiation and tip growth. Plant Cell. 2002;14(4):763–76.
42. Tan LSL, Nott DJ. Variational inference for generalized linear mixed models using partially noncentered parametrizations. Stat Sci. 2013;28:168–88.
43. Salimans T, Knowles DA. Fixed-form variational posterior approximation through stochastic linear regression. Bayesian Anal. 2013;8:837–82.
44. Syed S, Bouchard-Côté A, Deligiannidis G, Doucet A. Non-reversible parallel tempering: a scalable highly parallel MCMC scheme. J R Stat Soc Ser B Stat Methodol. 2022;84(2):321–50.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.