Bioinformatics, 35(21), 2019, 4327–4335
doi: 10.1093/bioinformatics/btz244
Advance Access Publication Date: 9 April 2019
Original Paper

OXFORD

Genetics and population analysis

A coordinate descent approach for sparse Bayesian learning in high dimensional QTL mapping and genome-wide association studies

Meiyue Wang and Shizhong Xu*

Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on April 2, 2018; revised on February 14, 2019; editorial decision on March 21, 2019; accepted on April 5, 2019

Abstract

Motivation: Genomic scanning approaches that detect one locus at a time are subject to many problems in genome-wide association studies and quantitative trait locus mapping. The problems include large matrix inversion, over-conservativeness for tests after Bonferroni correction and difficulty in evaluation of the total genetic contribution to a trait's variance. Targeting these problems, we take a further step and investigate a multiple locus model that detects all markers simultaneously in a single model.

Results: We developed a sparse Bayesian learning (SBL) method for quantitative trait locus mapping and genome-wide association studies. This new method adopts a coordinate descent algorithm to estimate parameters (marker effects) by updating one parameter at a time conditional on current values of all other parameters. It uses an L_2 type of penalty that allows the method to handle extremely large sample sizes ($>$100 000). Simulation studies show that SBL often has higher statistical powers and the simulated true loci are often detected with extremely small *P*-values, indicating that SBL is insensitive to stringent thresholds in significance testing.

Availability and implementation: An R package (sbl) is available on the comprehensive R archive network (CRAN) and https://github.com/MeiyueComputBio/sbl/tree/master/R%20packge.

Contact: shizhong.xu@ucr.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Most economically important traits in crops are quantitative in nature. Detecting marker-trait association is called quantitative trait locus (QTL) mapping when the population is created through line crossing experiments or genome-wide association studies (GWAS) if the population is a randomly collected sample of a natural population. In classical QTL mapping procedures, if a locus does not overlap with a marker, the numerical codes of the genotypes are inferred from two markers flanking the locus (defining an interval of the genome). Effects of QTL outside the interval are captured by co-factors included in the model, a method called composite interval mapping (CIM) (Jansen, 1993; Zeng, 1994). The unique feature of CIM is the inference of genotypes for pseudo markers. As the marker map becomes increasingly dense, it is redundant to infer genotypes of a pseudo marker and we may need to skip some markers because there are too many of them. With high marker density, the statistical technology used in GWAS has been adopted to QTL mapping. The co-factor selection step in CIM has been replaced by a polygenic effect modeled by a kinship matrix inferred from markers of the entire genome (Xu, 2013). We now have a unified linear mixed model as a powerful tool for both QTL mapping and GWAS because they share the same theoretical basis (Yu *et al.*, 2006).

One common feature of the mixed-model QTL mapping and GWAS is that both detect association of trait with one marker locus at a time until all loci are detected to complete the analysis. Technically, the above genome scanning approach uses a single QTL model. As a result, one often expects to see an island surrounding a high peak in a Manhattan plot which is caused by linkage

disequilibrium (LD). In contrast, the multiple locus model includes all markers in the same model with effects being estimated simultaneously. It drastically eliminates the island, leaving a single peak standing alone. This single peak is supposed to be better than an island because the signal is cleaner and stronger. This study particularly addresses this issue via Monte Carlo simulation studies. In the meantime, we developed a novel sparse Bayesian learning (SBL) method (Tipping, 2001) to perform multiple QTL linkage study and multiple locus GWAS by including all markers in a single model.

Multiple QTL linkage studies were invented two decades ago represented by the multiple interval mapping developed by Kao *et al.* (1999). Multiple locus GWAS is also available and mainly conducted via Markov chain Monte Carlo (MCMC). Current multiple marker models that have been applied to QTL mapping and GWAS include the least absolute shrinkage selection operator (LASSO) (Tibshirani, 1996), empirical Bayes (Xu, 2007), multi-locus mixed model (Segura *et al.*, 2012) and Bayes B (Meuwissen *et al.*, 2001). With the exception of LASSO, the other three methods are slow in terms of computational speed. LASSO is surprisingly fast for small and intermediate sample sizes (from ~200 to 5000) but can fail for large sample size with extremely high marker densities. In addition, the LASSO method implemented by GLMNET/R package does not provide a mechanism to calculate the standard errors of estimated effects and thus cannot perform statistical tests for markers. SBL (Tipping, 2001) is another method similar to LASSO but implemented with an L_2 penalty. Theoretically, SBL and empirical Bayes are the same. Both methods estimate prior variance of each effect and if the estimated variance is zero, the corresponding estimated effect is also zero (sparseness). When the tuning parameter favors sparseness, the prior variance has a very narrow legal space to move, beyond which the variance is set to zero. This explains why the L_2 penalty can also lead to sparseness.

Except multi-locus mixed model, none of the above methods is capable of dealing with large sample sizes. As the cost of DNA sequencing becomes increasingly cheaper, a large number of individuals are expected to be sequenced. A sample size of 5000 is considered typical for GWAS. In this study, we proposed a new SBL technique that can easily handle 100 000 individuals with a half million single nucleotide polymorphism (SNP) markers. Such a large sample size, although easily handled by a genome scanning approach (Lippert *et al.*, 2011; Loh *et al.*, 2015; Zhou and Stephens, 2012), has never been reported in QTL mapping and GWAS with a multiple locus model (Guan and Stephens, 2011; Segura *et al.*, 2012; Zhou *et al.*, 2013).

2 Materials and methods

2.1 Materials

2.1.1 QTL mapping in inbred rice

The population consists of 210 recombinant inbred lines (RIL) of rice derived from the cross of two elite inbred varieties, Zhenshan 97 and Minghui 63. The hybrid of the two varieties, Shanyou 63, underwent nine generations of selfing via single-seed descent to generate the 210 RILs, which were evaluated in 1997, 1998 and 1999 in two locations at the Experimental Station of Huazhong Agricultural University in Wuhan, China. We analyzed two agronomic traits: 1000 grain weight (KGW) and yield per plant (YD). The genotypes consist of 1619 bins inferred from 270 820 SNPs across 12 chromosomes of the rice genome. A bin is a synthetic locus covering all markers that share the same segregation pattern in a complete LD block. We took the average value of a trait across the four replicates as the original phenotypic value for each trait. Details of the rice experiment are provided in the original publication (Yu *et al.*, 2011).

2.1.2 GWAS in hybrid rice

The hybrid population analyzed in this study consists of 1495 hybrid rice varieties derived from *indica*×*indica* (1439), *indica*×*japonica* (18) and *japonica*×*japonica* (38) crosses. The original 96-bp paired-end sequencing reads and phenotype dataset were obtained from Huang *et al.* (2015). The genotype datasets were downloaded from the Rice Haplotype Map Project website (http://www.ncgr.ac.cn/RiceHap4). We realigned reads against the reference genome of *japonica* Nipponbare (MSU Rice Genome Annotation Project Release 7) and performed SNP calling. After SNP filtration of missing rate <5% and minor allele frequency <1%, 182 010 SNPs of the 1495 hybrid varieties were randomly selected across the genome for association analysis. Each hybrid variety was planted in two experimental fields in Hangzhou, China (subtropical and long-day condition) and Sanya, China (tropical and short-day condition). In this study, we used the average of phenotypic values collected from the two locations as the original response variable.

2.2 Hierarchical linear mixed model

Let y be a vector of phenotypic values of a quantitative trait collected from n individuals. Define Z_{jk} as a genotype indicator variable of individual j at marker k with three values, 1, 0 and −1, representing the three possible genotypes of locus k. The linear mixed model for y is

$$y = \sum_{l=1}^{q} X_l\beta_l + \sum_{k=1}^{m} Z_k\gamma_k + \varepsilon \quad (1)$$

where X_l and β_l represent the design matrix and the effect for the lth fixed effect (non-genetic), Z_k is a genotype indicator for marker k, γ_k is the effect of this marker and ε is the residual error with an assumed $\varepsilon \sim N(0, I\sigma^2)$ distribution. The marker effect γ_k is treated as a random variable with an assumed $N(0, \phi_k^2)$ distribution, where ϕ_k^2 is a prior variance that must be estimated from the data. To control the sparseness of the model, we assign a hierarchical prior distribution to ϕ_k^2, an inverse Chi-squared distribution $p(\phi_k^2) \propto (\phi_k^2)^{-(\tau+2)/2}$, where τ (degree of freedom) is a hyper parameter. Each of the fixed effects (β_l) and the residual variance (σ^2) has a default uniform prior (also called uninformative prior).

2.3 Conditional posterior mode estimation of marker effects

We adopt a coordinate descent algorithm (Ortega and Rheinboldt, 1970) to estimate one parameter at a time conditional on values of other parameters. Let us define a new linear mixed model by

$$y_l = y - \sum_{l' \neq l}^{q} X_{l'}\beta_{l'} - \sum_{k=1}^{m} Z_k\gamma_k = X_l\beta_l + \varepsilon \quad (2)$$

which can be interpreted as the phenotypic values adjusted by all other effects except $X_l\beta_l$. Such an adjusted phenotypic vector allows us to obtain the conditional posterior mode estimate of β_l using $y_l = X_l\beta_l + \varepsilon$. The simple least squares estimate (posterior mode) of β_l conditional on all other effects is $\hat{\beta}_l = (X_l^T X_l)^{-1}(X_l^T y_l)$ for $l = 1, \ldots, q$. Note that this least squares estimation only involves the inverse of a scalar. When we say conditional on some parameters, these parameters are treated as known constants. For example, y_l is a function of all parameters except β_l, but all other parameters are assumed to be known and thus the conditional estimate of β_l has a very simple form (Ortega and Rheinboldt, 1970).

Since γ_k is a random effect, its estimate is called the best linear unbiased prediction (BLUP). The conditional BLUP of γ_k given all other effects is obtained using the following linear random model

$$y_k = y - \sum_{l=1}^{q} X_l\beta_l - \sum_{k'\neq k}^{m} Z_{k'}\gamma_{k'} = Z_k\gamma_k + \varepsilon \tag{3}$$

where y_k is the phenotypic value adjusted by all effects except $Z_k\gamma_k$. The coordinate descent algorithm is an iterative approach where only one parameter is studied at a time conditional on all other parameters. In other words, γ_k is the only parameter in model (3) because all β's and other γ's with subscripts not equal to k are treated as knowns. For this simple model, $y_k = Z_k\gamma_k + \varepsilon$, the conditional variance of y_k given γ_k is $\mathrm{var}(y_k|\gamma_k) = \mathrm{var}(\varepsilon) = I\sigma^2$, the residual variance. The following equations are derived based on this assumption. In the next section, we define the variance of y_k conditional on all other parameters except γ_k by $\mathrm{var}(y_k) = Z_kZ_k^T\phi_k^2 + I\sigma^2$. The BLUP of γ_k is simplified using the Woodbury matrix identity (Woodbury, 1950),

$$\hat{\gamma}_k = \lambda_k Z_k^T y_k - \frac{\lambda_k^2 (Z_k^T Z_k)(Z_k^T y_k)}{\lambda_k Z_k^T Z_k + 1} \tag{4}$$

where $\lambda_k = \phi_k^2/\sigma^2$ is the ratio of variance components. The variance of $\hat{\gamma}_k$ is

$$\mathrm{var}(\hat{\gamma}_k|y_k) = \left[\lambda_k - \lambda_k^2\left(Z_k^TZ_k - \lambda_k\frac{(Z_k^TZ_k)^2}{\lambda_kZ_k^TZ_k+1}\right)\right]\sigma^2. \tag{5}$$

The estimated residual variance conditional on all model effects is

$$\sigma^2 = \frac{(y - \sum_{l=1}^{q} X_l\hat{\beta}_l - \sum_{k=1}^{m} Z_k\hat{\gamma}_k)^T(y - \sum_{l=1}^{q} X_l\hat{\beta}_l - \sum_{k=1}^{m} Z_k\hat{\gamma}_k)}{n - q - m_0} \tag{6}$$

where

$$m_0 = \sum_{k=1}^{m}\lambda_k\left(Z_k^TZ_k - \frac{\lambda_kZ_k^TZ_kZ_k^TZ_k}{\lambda_kZ_k^TZ_k+1}\right) \tag{7}$$

is the effective number of markers (Tipping, 2001; Xu, 2013). Derivations of the BLUP and their variances along with m_0 are presented in Supplementary File S1.

The BLUP estimates of marker effects depend on $\lambda_k = \phi_k^2/\sigma^2$. When ϕ_k^2 is replaced by the estimated value, the estimate is no longer BLUP; it is called the empirical Bayes estimate. Therefore, we need to estimate ϕ_k^2 also from the dataset. The SBL of Tipping (2001) does not involve β_l and γ_k during the iteration process; rather, it deals with ϕ_k^2 (the prior variance of γ_k) in the process of optimization.

2.4 Conditional posterior mode estimation of marker variances

We now derive a simple solution for ϕ_k^2. The variance of the random model given in Equation (3) is $\mathrm{var}(y_k) = Z_kZ_k^T\phi_k^2 + I\sigma^2$. The logarithmic posterior probability of ϕ_k^2 after incorporating the hyper parameter is

$$L(\phi_k^2) = -\frac{1}{2}\ln(Z_k^TZ_k\phi_k^2/\sigma^2+1) + \frac{\phi_k^2y_k^TZ_kZ_k^Ty_k}{2\sigma^4(Z_k^TZ_k\phi_k^2/\sigma^2+1)} - \frac{\tau+2}{2}\ln(\phi_k^2) \tag{8}$$

where terms irrelevant to ϕ_k^2 have been ignored. This logarithmic posterior is conditional on all $\gamma_{k'}$ for $k' \neq k$, while the posterior of Tipping (2001) is conditional on all $\phi_{k'}^2$ for $k' \neq k$. Such a modification results in the following explicit solution for ϕ_k^2. Let $s_k = Z_k^TZ_k/\sigma^2$ and $h_k = Z_k^Ty_k/\sigma^2$, we get

$$L(\phi_k^2) = -\frac{1}{2}\ln(s_k\phi_k^2+1) + \frac{\phi_k^2h_k^2}{2(s_k\phi_k^2+1)} - \frac{\tau+2}{2}\ln(\phi_k^2). \tag{9}$$

Obviously, the global solution is $\phi_k^2 = 0$, but we need a local solution in this case. Setting $\partial L(\phi_k^2)/\partial\phi_k^2 = 0$ leads to

$$-(\tau+3)s_k^2(\phi_k^2)^2 - [(2\tau+5)s_k - h_k^2]\phi_k^2 - (\tau+2) = 0 \tag{10}$$

which is a quadratic function of ϕ_k^2 with a local solution (the largest positive solution) equal to $\phi_k^2 = [-b - (b^2-4ac)^{1/2}]/(2a)$, where $a = -(\tau+3)s_k^2$, $b = h_k^2 - (2\tau+5)s_k$ and $c = -(\tau+2)$. Detailed derivation of Equation (10) is presented in Supplementary File S1. Whenever a solution is negative or illegal, we should take the global solution $\phi_k^2 = 0$, leading to sparseness of the model. Note that we can use τ to control the model sparseness, where $-2 \leq \tau \leq 0$. When we set $\tau = -2$, the solution is $\phi_k^2 = (h_k^2 - s_k)/s_k^2$, which represents the least sparseness. Again, if $h_k^2 < s_k$, ϕ_k^2 is set to zero. The sparseness will increase as τ increases. We will discuss how to set an optimal value of τ in the result section later. The explicit solution for each ϕ_k^2 is crucial for the high computational efficiency of our new SBL. This approach has been applied to variable selection and it is called iterative conditional mode algorithm (Pungpapong *et al.*, 2015).

2.5 Summary of the coordinate descent algorithm

The iteration process is summarized as follows.

Step 1: Initialize the following variables, $\beta_l = \gamma_k = 0$ and $\sigma^2 = 1$.

Step 2: Update one β_l at a time until all β_l's values have been updated.

Step 3: Estimate ϕ_k^2 and thus $\lambda_k = \phi_k^2/\sigma^2$ for all $k = 1, \ldots, m$.

Step 4: Update $\gamma_k$using BLUP given in Equation (4) for all $k = 1, \ldots, m$.

Step 5: Update σ^2 based on updated $\sum_{l=1}^{q} X_l\beta_l$ and $\sum_{k=1}^{m} Z_k\gamma_k$.

Step 6: Repeat steps (2)–(5) until each parameter converges to a constant.

One important property of the algorithm is that $\sum_{l=1}^{q} X_l\beta_l$ and $\sum_{k=1}^{m} Z_k\gamma_k$ are updated instantly when an effect (β_l or γ_k) is estimated (instead of waiting until all effects are estimated). Theoretical computing time complexity of SBL is $O(q^3 + nq^2 + mnt_S + nqt_S)$ where t_S is the number of iterations required for the iterations to converge. The computational cost can be simplified to $O(nq^2 + mnt_S)$ since both m and n are substantially larger than q.

2.6 Simulation

We simulated $n = 500$ and $n = 1000$ individuals of an F_2 family generated from the cross of two inbred lines. The total number of markers for the entire genome (two chromosomes) was $m = 2000$. The first chromosome contained 20 QTLs with effects and positions shown in Figure 1a as well as in Supplementary Table S1. The second chromosome contained no QTL and this 'empty' chromosome was used to control Type 1 error in a separate QTL mapping study. The genetic map used in the simulation is provided in Supplementary File S2. Phenotypes of the n individuals were generated using $y = \beta_0 + \sum_{k=1}^{m} X_k\beta_k + e$, where $\beta_0 = 100$ is the intercept, X_k is the numeric genotype indicator of the marker k, β_k is the effect assigned to that marker and e is the residual error vector following an $N(0, I\sigma^2)$ distribution with $\sigma^2 = 10$ and $\sigma^2 = 20$, respectively. Details of individual marker variances and covariance between the

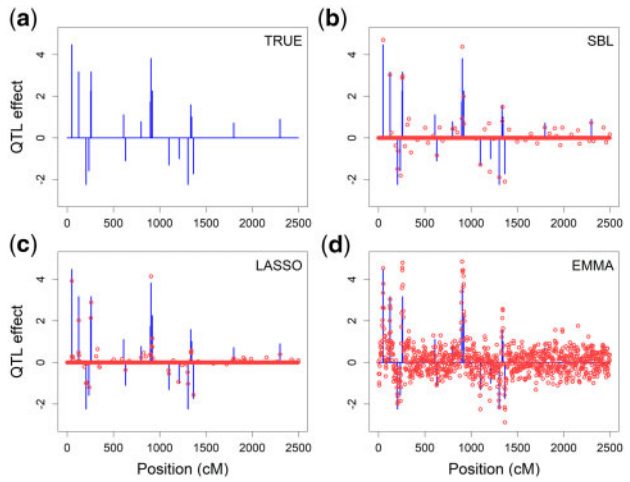

Fig. 1. Estimated effects of 1000 simulated markers on the first chromosome when the sample size is $n = 500$ and the residual error variance is $\sigma^2 = 10$. (**a**) True QTL effects (blue needles); (**b**) estimated marker effects from SBL (red dots); (**c**) estimated marker effects from LASSO (red dots) and (**d**) estimated marker effects from EMMA (red dots). (Color version of this figure is available at *Bioinformatics* online.)

simulated QTLs are given in Supplementary File S3. When $\sigma^2 = 10$, the heritability of the simulated trait is $H = \sigma_G^2/(\sigma_G^2 + \sigma^2) = 0.8559$. Contribution from each QTL varies from 0.36–14.39%. When $\sigma^2 = 20$, the heritability is $H = 0.7482$ and the contribution from each QTL varies from 0.32 to 12.58% (Supplementary Table S1). The two sample sizes combined with the two residual error variances produces a combination of four experimental setups: (1) $(n, \sigma^2) = (500, 10)$, (2) $(n, \sigma^2) = (500, 20)$, (3) $(n, \sigma^2) = (1000, 10)$ and (4) $(n, \sigma^2) = (1000, 20)$. Under each experimental setup, the simulation was replicated 100 times. Theoretical background of the simulation experiments is presented in Supplementary File S4 and a sample genotype and phenotype data with experimental setup $n = 500$ and $\sigma^2 = 10$ are provided in Supplementary Files S5 and S6.

In consideration of the negative interference from LD pattern and other intrinsic hidden structure of a real population, we randomly extracted $m = 5,\ 10$ and $50k$ (m is the number of markers and $k = 1000$) subset SNPs from the hybrid rice genome (Huang *et al.*, 2015) keeping the original sample size ($n = 1495$). The same 20 QTLs introduced earlier were randomly assigned to the entire genome. Phenotypes of the 1495 hybrids were generated using equation $y = \beta_0 + \sum_{k=1}^{m} X_k\beta_k + e$ described previously with only $e \sim N(0,\ 10)$.

To understand how the number of causal QTL affects the performance of each method, we fixed the total number of markers at $m = 10k$ and varied the number of causal SNPs at three levels, $m_{\mathrm{QTL}} = 40, 100, 160$. The positions of the causal SNPs were randomly assigned on the genome. Under each level of m_{QTL}, the simulation was replicated 100 times. For one of the simulated datasets, we permuted the phenotypes 1000 times to generate a permuted sample of test statistic, from which an empirical threshold was calculated.

2.7 Statistical power and Type 1 error

Wald test statistic, $W_k = \hat{\gamma}_k^2/\mathrm{var}(\hat{\gamma}_k|y_k)$, was used to test the null hypothesis $H_0 : \gamma_k = 0$. Under the null model, W_k follows approximately a Chi-squared distribution with one degree of freedom. Therefore, the P-value of each marker was calculated as $p_k = 1 - \Pr(\chi_1^2 \leq W_k)$. By approximation, we mean that the distribution of the random effect Wald test is not known. When the standard error of the estimated effect is small, we can treat the random effect as a 'fixed effect' and thus the Wald test can be approximated by the Chi-squared distribution. In the significance test for the LASSO method, Lockhart *et al.* (2014) found that the test statistic follows an Exp(1) distribution. The nominal probability with Bonferroni correction for multiple tests, $0.05/m$, is used as the threshold. The power for each simulation experiment is defined as the proportion of detected QTLs over the total number of simulated QTLs. Because of LD, markers nearby each simulated QTL are often detected. Therefore, we reserved a three-marker window around each QTL. In total, there are m_{QTL} QTL windows covering $3m_{\mathrm{QTL}}$ markers. If any markers of the triplet were detected, we counted the triplet as one positive detection. If a marker outside the QTL windows was detected, it was counted as one false positive detection. The Type 1 error is defined as the proportion of false positives over the $m - 3m_{\mathrm{QTL}}$ markers. The false discovery rate (FDR) is defined as the proportion of false positives among all detected markers.

2.8 Alternative thresholds of test statistics

In the simulation studies, we also evaluated powers and Type 1 errors empirically using different thresholds of the Wald test. In the first approach, we picked up the maximum Wald test over all 1000 markers from the second chromosome for each of the 100 replicated simulation experiments. We then ranked the 100 maximum Wald tests in ascendant order and chose the 95th percentile as the threshold value (called Threshold-A). In the second approach, we included all markers in the non-QTL windows of the first chromosome and all markers in the second chromosome in the pool (a total of 1940 markers). For each of the 100 replicated simulations, we chose the maximum Wald test over all these markers. The 95th percentile of the 100 maximum Wald tests was used as the threshold value (called Threshold-B).

3 Results

3.1 Tuning the hyper parameter τ of the SBL method

The inverse Chi-squared distribution assigned to ϕ_k^2 is $p(\phi_k^2) \propto (\phi_k^2)^{-(\tau+2)/2}$, where $\tau \geq -2$. We used the sample data provided by the GLMNET/R package (Friedman *et al.*, 2010) (the dataset contain 100 observations and 20 variables) to demonstrate how model sparseness changes with different values of τ. Setting $\tau = 61.25$, SBL selects the first non-zero coefficient and more coefficients appear when τ is further decreased (Supplementary Fig. S1a). In the interval of $-2 \leq \tau \leq 0$, most coefficients are estimated nearly constantly. Only one more coefficient emerges within that interval of τ and the estimated value of this extra coefficient is close to 0. Thus $-2 \leq \tau \leq 0$ is an optimal interval for tuning τ. We extended the original simulated experiment of $(n, m, \sigma^2) = (1000, 2000, 10)$ to different levels of m ($m = 2, 10, 50, 100, 300, 500k$). We set τ as a sequence of values within $[-2, 0]$ to fit the model and evaluated the model predictability using the leave-one-out cross validation on these simulation experiments (Xu, 2017). The optimal tuning parameter takes the value (between -2 and 0) that produce the highest predictability. Although there is no direct connection between the predictability and the statistical power of detection, using predictability as the criterion to tune parameters is commonly practiced in model selection (Tibshirani, 1996). When $m \leq 10k$, the predictability of SBL decreases as τ increases; the predictability is stable when the number of markers is intermediate $50k \leq m \leq 100k$; when $m \geq 300$k, larger τ results in higher predictability (Supplementary Fig. S2).

Based on our experience, fixing the tuning parameter at $\tau = -1$ is an eclectic solution to handle general cases without performing cross validation. Therefore, we fixed $\tau = -1$ in all subsequent simulation studies.

3.2 Simulation studies

In the simulation study (the F2 family), we compared the new SBL method with efficient mixed-model association (EMMA) (Kang *et al.*, 2008) and LASSO (Tibshirani, 1996). We used the GLMNET/R package (Friedman *et al.*, 2010) to implement the LASSO method and estimated the standard errors for estimated marker effects via two approaches proposed by Ithnin *et al.* (2017). One (LASSO-A) is based on the bootstrap method (Efron and Tibshirani, 1994) and the other (LASSO-B) is an approximate method incorporating Henderson's mixed-model equation (Henderson, 1975). For each method, we used four criteria to test the significance of a marker: (i) comparing the marker *P*-value with $0.05/m = 2.5\text{E}-5$ (after Bonferroni correction), (ii) controlling FDR at nominal probability $q = 0.2$ (Benjamini and Hochberg, 1995), (iii) comparing the Wald test with Threshold-A and (iv) comparing the Wald test with Threshold-B. Criteria (iii) and (iv) are permutation-based methods.

Figure 1 shows the estimated marker effects plotted against genome locations of the markers under $n = 500$ and $\sigma^2 = 10$. Estimates of SBL (Fig. 1b) and LASSO (Fig. 1c) are much closer to the true effects (Fig. 1a), both being sparse in a sense that most markers have zero estimated effects. The EMMA method (Fig. 1d), however, is not sparse and has a substantially noisy background. Similar plots under other experimental setups are shown in Supplementary Figures S3–S5.

Figure 2 shows the statistical powers and Type 1 errors of the four methods under all four experimental setups. Overall, SBL has the highest power followed by EMMA, LASSO-B and then LASSO-A. The Bonferroni corrected threshold may be too stringent for LASSO. The Type 1 error of all four methods is well controlled below 0.008, and the LASSO method even has a Type 1 error =0. Supplementary Figure S6 shows the power and the FDR of the four methods with 0.2 as the controlled FDR. Statistical powers of all four methods are increased but the FDR of SBL and EMMA are not well controlled as expected. When the alternative thresholds (Threshold-A and Threshold-B) are used, the powers and Type 1 errors for the four methods are shown in Supplementary Figures S7 and S8, respectively. For Threshold-A, EMMA often has higher power but is also associated with much higher Type 1 errors. The three multiple locus methods have virtually zero Type 1 error. For Threshold-B, the three multiple locus models have higher powers than EMMA, and the Type 1 errors of all methods are extremely low.

The best way to represent the nature of the four methods is through the receiver operating characteristic (ROC) curves, which are illustrated in Figure 3, where the powers of all methods are compared at the same level of Type 1 error. When the Type 1 error is small, SBL always shows higher power than the other methods in three of the four experimental setups. The exception occurs in the situation where $n = 500$ and $\sigma^2 = 20$. The EMMA method is the least efficient of the four methods compared. Supplementary Figure S9 shows the ROC curves comparing powers at the same level of FDR, which has the same trend as Figure 3.

The ROC curves are much the same for the multiple locus models in the simulation studies under the four different experimental setups. We simulated more data with large n and m to further compare the computing time. We ran both methods on the high-performance computing Linux cluster hosted by the Bioinformatics Facility at the University of California Riverside. LASSO-A (the bootstrap version) is still too much for the high-performance computing cluster system, thus we only compared SBL with LASSO-B. We investigated 13 cases with various n and m combinations as shown in Supplementary Table S2; it also shows the running time with one CPU core without parallel computing. When sample size is intermediate, and the number of markers is large, LASSO-B

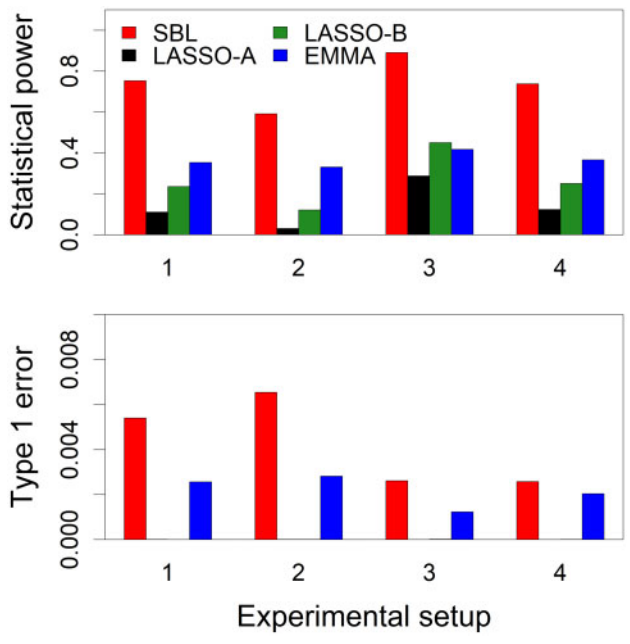

Fig. 2. Statistical powers (upper panel) and Type 1 errors (lower panel) obtained at $p = 0.05/m = 2.5E-5$ drawn from 100 replicated simulation experiments of four methods under four different experimental setups. The four experimental setups are four combinations of sample size and residual error variance: (i) $n = 500$ and $\sigma^2 = 10$; (ii) $n = 500$ and $\sigma^2 = 20$; (iii) $n = 1000$ and $\sigma^2 = 10$ and (iv) $n = 1000$ and $\sigma^2 = 20$

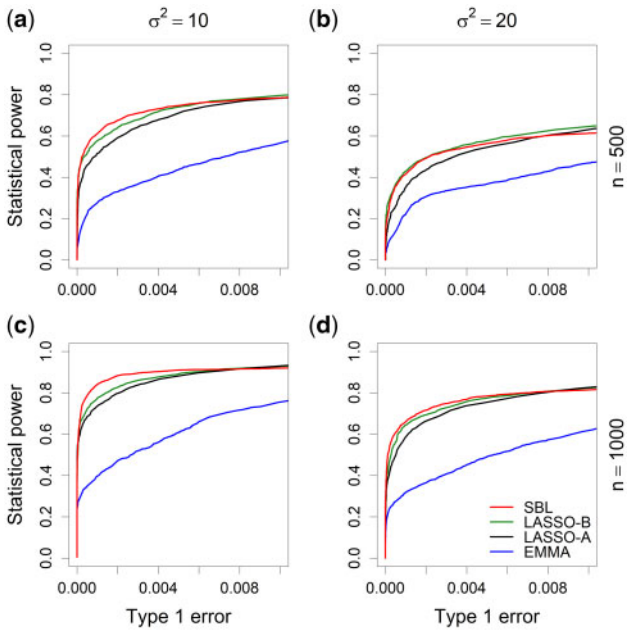

Fig. 3. ROC curves of four methods drawn from 100 replicated simulations under four experimental setups: (**a**) $n = 500$ and $\sigma^2 = 10$; (**b**) $n = 500$ and $\sigma^2 = 20$; (**c**) $n = 1000$ and $\sigma^2 = 10$ and (**d**) $n = 1000$ and $\sigma^2 = 20$

outcompetes SBL in terms of computational speed. However, LASSO-B failed to handle sample size $100k$ when $m \geq 50k$. When the sample size is $5k$ but the number of markers is $500k$, LASSO also failed due to memory problem. However, under the situations where LASSO failed, SBL remained feasible and completed the analysis within a reasonable time frame.

The second simulation study was performed using the hybrid rice genotypes. We were able to apply the LD score regression technique (Bulik-Sullivan *et al.*, 2015) on these datasets so that the SBL can be compared with BOLT-LMM (Loh *et al.*, 2015). We also compared the new method with another multiple locus model called BSLMM (Zhou *et al.*, 2013). We converted our data into the PLINK format (Purcell *et al.*, 2007) because the BSLMM program takes input data with that format. The BSLMM method is a Bayesian method implemented via the MCMC sampling algorithm. The total length of the chain was 50 000 iterations, the burn-in period was 20 000 iterations and thereafter the chain was thinned by deleting 9 out of 10 iterations. We extracted $m = (5, 10, 50k)$ markers from the hybrid rice genome and assigned $m_{\mathrm{QTL}} = 100$ causal markers (QTLs) randomly on the genome. Because the BSLMM program provides a posterior inclusion probability (PIP) to each marker rather than the *P*-value, we performed 1000 permutations to find the empirical threshold of PIP for valid comparison among SBL, BOLT-LMM and BSLMM. The maximum PIPs from the 1000 permutations were ranked in ascending order and the 95th percentile was used as cutoff criterion for BSLMM. For SBL and BOLT-LMM, we ranked 1000 minimum *P*-values and chose the fifth percentile as the significance criterion. Under the empirical thresholds obtained by the same approach, SBL appears to have lower powers than BOLT-LMM and BSLMM but the FDRs are well controlled, while BOLT-LMM and BSLMM do not control the FDR properly (Supplementary Table S3). We also fixed the number of markers at $m = 10k$ and varied $m_{\mathrm{QTL}} = (40, 100, 160)$ to evaluate the effect of causal QTL number on the power. Supplementary Table S4 shows the power comparison of the three methods. Again, although BOLT-LMM and BSLMM have higher power than SBL, they do not control the FDR properly.

Since there are big differences in FDR for the three methods, we then compared the powers of the three methods under the same false positive sequence. Supplementary Figure S10 shows the ROC curve comparison when $m = (5, 10, 50k)$ under the same number of causal QTLs ($m_{\mathrm{QTL}} = 100$). In the cases of $m = (5k, 10k)$, SBL and BSLMM are more powerful than BOLT-LMM. When $m = 50k$, the statistical power of BOLT-LMM surpasses the powers of SBL and BSLMM after the Type 1 error rate is >0.003 and 0.004, respectively. When we fixed $m = 10k$ but varied $m_{\mathrm{QTL}} = (40, 100, 160)$, it appears that larger number of causal QTLs favors the SBL method because the power of SBL increases as more QTLs are included, and BOLT-LMM is the least powerful method in this case (Supplementary Fig. S11). We believe that the overall high power of SBL and BSLMM over BOLT-LMM is due to the multiple locus model and the sparseness of the model implemented by both methods.

Computing time of SBL, LASSOs, EMMA, BOLT-LMM and BSLMM with programs' default setting on the second simulation study is provided in Supplementary Table S5. SBL and LASSO-B are very fast, taking about 10 min for populations with 1500 individuals and 200 000 markers. However, BSLMM is very slow due to the length of MCMC samplings. We also compared the maximum memory usages of various methods under comparison (Supplementary Fig. S12). As expected, single locus methods (EMMA and BOLT-LMM) use less memory compared to the multiple locus methods (SBL, LASSO-A, LASSO-B and BSLMM). Among the four multiple locus methods, SBL requires the least memory. The BSLMM appears to be the most powerful method, but it is also the slowest method.

3.3 Mapping QTL for KGW in rice with a RIL population

Four methods (SBL, LASSO-A, LASSO-B and EMMA) were used to map QTL for KGW in the RIL population of rice. $\tau = 0$ was used in SBL for sparse model fitting. Since we took the average value of a trait collected from four replicates as the phenotypic value of each trait, the intercept was the only fixed effect included in the model. SBL and EMMA detected a known QTL on chromosome 3 (*GS3*) and another known QTL on chromosome 5 (*GW5/qSW5*). *GW5/qSW5* was also identified by LASSO-B (Table 1). The $-\log_{10}(p)$ test statistics were plotted against the genome location (Supplementary Fig. S13). No significant QTL was detected by LASSO-A, perhaps due to the very stringent criterion after Bonferroni correction. In this plot, a peak in LASSO-A is found corresponding to *GW5/qSW5* but that peak does not pass the threshold. In addition to the two known QTLs, SBL detected nine additional QTLs (Table 1). Among the 11 QTLs detected by SBL, *kgw1.6*, *kgw1.33*, *kgw3.16*, *kgw3.28*, *kgw5.5* and *kgw9.19* are consistent with the QTLs detected by Yu *et al.* (2011). The *kgw3.16* QTL is mapped to a region defined by an interval (15.597–16.914 Mb) on chromosome 3 that contains *GS3*, which is a well-studied QTL to control grain length (GL) (Fan *et al.*, 2006). All three methods identified *kgw5.5* that is accurately mapped to the interval defined by 4.776–5.376 Mb on chromosome 5, which contains gene *GW5/qSW5* that is known to control grain width (Weng *et al.*, 2008). The BOLT-LMM method was not used here because we do not have the reference LD information of the RIL population required by the method.

We further investigated the associations between markers and the yield trait. The SBL method detected five QTLs, but none of the remaining three methods detected any QTLs (Supplementary Figure S14 and Table S6). Among these five QTLs, *yd1.33* overlaps with *kgw1.33*. A significant QTL *yd7.8* is mapped to the interval bracketed by 8.407–8.756 Mb on chromosome 7 with $p = 3.33E-15$ and the genetic position is 54.008 cM. The physical position of *yd7.8* corresponds to the peak of EMMA, though the EMMA peak does not pass the threshold. This QTL (*yd7.8*) was also identified by Yu (Yu *et al.*, 2011) and the mapping interval contains *Ghd7* (Xue *et al.*, 2008), a major QTL to control the number of grains per

Table 1. Significant QTLs identified by SBL, LASSO-B and EMMA for KGW of rice from the RIL population

Method	QTL name	Chromosome	Interval (Mb)	Position (cM)	*P*-value
SBL	*kgw1.6*	1	6.232–6.272	36.06	3.77E-22
	kgw1.33	1	32.718–33.285	145.255	7.85E-16
	kgw3.16	3	15.597–16.914	93.752	1.28E-37
	kgw3.28	3	28.511–28.598	131.88	2.88E-17
	kgw5.5	5	4.776–5.376	29.709	3.95E-52
	kgw5.25	5	25.281–25.902	102.386	4.72E-06
	kgw6.1	6	1.366–1.514	5.819	3.15E-06
	kgw6.12	6	12.49–13.724	68.453	2.71E-20
	kgw7.8	7	7.595–8.407	52.253	3.19E-07
	kgw9.19	9	19.805–20.063	86.333	1.87E-21
	kgw11.9	11	9.031–9.294	53.036	6.89E-12
LASSO-B	*kgw5.5*	5	4.776–5.376	29.709	9.02E-31
EMMA	*kgw3.16*	3	15.597–16.914	93.752	1.57E-05
	kgw5.5	5	4.776–5.376	29.709	1.04E-12

panicle as well as a pleiotropic QTL that affects yield, heading date and plant height.

3.4 GWAS for GL in hybrid rice

All six methods of GWAS (SBL, LASSO-A, LASSO-B, EMMA, BOLT-LMM and BSLMM) were compared for the hybrid rice population. We set the hyper parameter $\tau = 0$ in the SBL method to fit a sparser model for this dense marker dataset. The Manhattan plots for GL are shown in Figure 4 for all methods, where we arbitrarily truncated any markers with $-\log_{10}(p) > 15$ to $-\log_{10}(p) = 15$ to improve the visibility of the plots. For BSLMM, the PIP of each SNP instead of $-\log_{10}(p)$ was plotted against the genome. A total of 123 markers were significant, where 15 of them were detected by SBL (Supplementary File S7), two by LASSO-A, zero by LASSO-B, 94 by EMMA, six by BOLT-LMM and six by BSLMM. We matched the significant SNPs within 100 kb (both upstream and downstream) of known genes that have been cloned and experimentally validated to control GL. Only one known gene, *GS3* (Fan *et al.*, 2006), had been detected by SBL, EMMA and BOLT-LMM with Bonferroni corrected threshold $p = 2.75\text{E} - 7$ and by BSLMM with a predetermined nominal probability of $1 - \alpha = 0.95$.

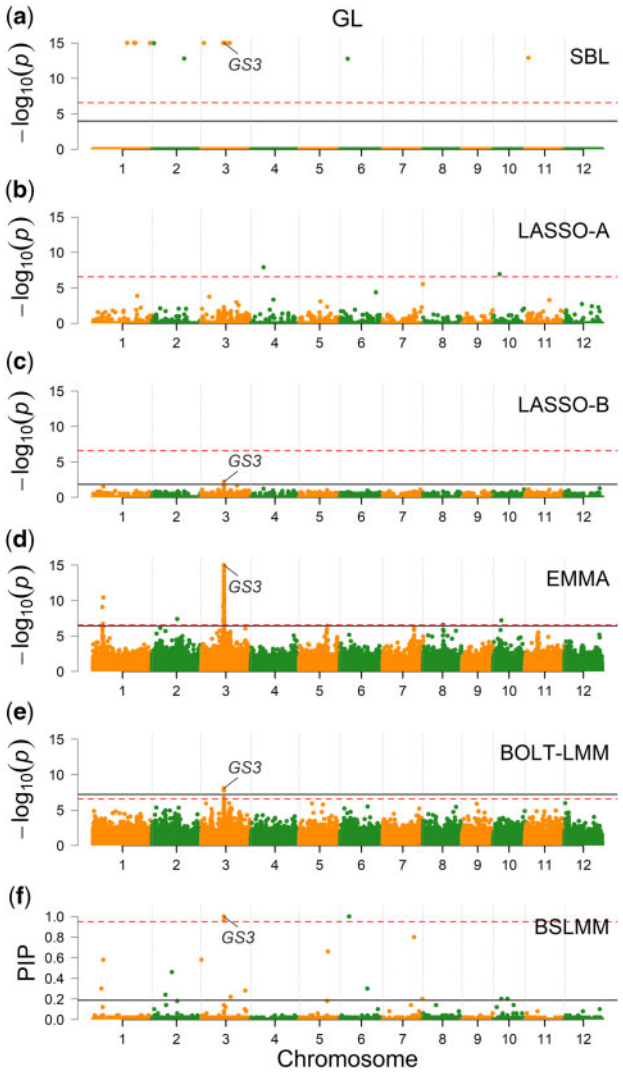

Fig. 4. Manhattan plots for GL of the hybrid rice varieties obtained from six methods: (**a**) SBL, (**b**) LASSO-A, (**c**) LASSO-B, (**d**) EMMA, (**e**) BOLT-LMM and (**f**) BSLMM. The red dashed lines indicate the genome-wide threshold $-\log_{10}(0.05/182010) = 6.56$ for SBL, LASSO-A, LASSO-B, EMMA and BOLT-LMM, while the threshold of PIP is $1 - \alpha = 0.95$ for BSLMM. The solid gray lines indicate the empirical threshold generated from 1000 permutations for SBL ($-\log_{10}(0.000106) = 3.97$), LASSO-B ($-\log_{10}(0.01444) = 1.84$), EMMA ($-\log_{10}(3.86E-7) = 6.41$), BOLT-LMM ($-\log_{10}(6.4E-8) = 7.19$) and BSLMM ($1 - \alpha = 0.1866667$). The annotated SNPs overlap with a known gene, *GS3*, that controls GL. (Color version of this figure is available at *Bioinformatics* online.)

We further permuted the phenotypes of GL 1000 times to construct a null distribution of the statistic to find the empirical threshold for each method. The minimum *P*-value from 1000 permutations were ranked in ascending order and the fifth percentile of the 1000 minimum *P*-values was used as the empirical threshold. For BSLMM, the maximum PIP from the 1000 permutations was ranked and we chose the 95th percentile as the cutoff criterion. The new criterion for SBL was $p = 0.000106$ and the same 15 SNPs were detected, including the known gene *GS3*. We did not conduct permutations for LASSO-A because the bootstrap step along with the permutations would take excessively long time to finish. LASSO-B had a new genome-wide threshold of $p = 0.01444$ and a significant SNP was identified with this new threshold. This SNP overlapped with the cloned QTL *GS3*. EMMA detected five additional SNPs with the new threshold $p = 3.86E-7$ and none of the five SNPs matched any known gene for GL. The new threshold of BOLT-LMM happened to be $p = 6.40E-8$ (lower than the previous threshold). With the new threshold, BOLT-LMM identified four SNPs in total, all overlapping with the known gene *GS3* (within ±100 kb). The empirical criterion for BSLMM had dropped to $1 - \alpha = 0.1866667$ and 13 additional SNPs were detected with this threshold. However, none of these SNPs overlapped with any known genes that control GL. The large difference between the typical significance level and the empirical threshold indicates that $1 - \alpha = 0.95$ maybe too stringent for BSLMM.

We also detected associations of markers with panicle number (PN) and panicle length (PL) for the rice hybrid population. For PN, the SBL method detected 19 associated markers, LASSO-A detected one marker, EMMA detected two markers, BOLT-LMM detected 19 markers (Supplementary Fig. S15). No markers were identified by LASSO-B and BSLMM for PN. Among the 19 markers detected by SBL, two of them overlapped with identified genes *Cga1* (Hudson *et al.*, 2013) and *OscpSRP43* (Lv *et al.*, 2015), respectively (Supplementary File S7). One out of the 19 markers identified by BOLT-LMM matched a known gene associated with PN, called *FUWA* (Chen *et al.*, 2015). The SBL method identified 18 significant SNPs associated with PL (Supplementary Fig. S16). One known gene was detected, *OsDET1* (Zang *et al.*, 2016) (Supplementary File S7). LASSO-A only detected one marker without matching any known gene and no marker was identified by LASSO-B. Among 81 SNPs detected by EMMA, four SNPs matched known genes that affect the PL trait, including *CCP1* (Yan *et al.*, 2015), *AVB* (Ma *et al.*, 2017), *RLS3* (Lin *et al.*, 2016) and *MRG702* (Jin *et al.*, 2015). BOLT-LMM and BSLMM detected five and one significant markers associated with PL, respectively, and none of them overlapped with any known genes.

4 Discussion

Statistical methods for QTL mapping and GWAS have long been studied separately. The two technologies are designed for dealing with different types of populations and marker densities. As the rapid advancement of molecular technology, genotyping high density markers has been feasible for almost all species and marker density can be very high even for linkage mapping populations. From

this point of view, the two areas have been merged. In QTL mapping, genetic background is captured via CIM by fitting co-factors (Jansen, 1993; Zeng, 1994), while in GWAS, the polygenic background is modeled via a kinship matrix inferred from genome-wide markers (Zhou and Stephens, 2012). With the multiple locus model, genetic background control is no longer needed because all markers are fitted to the model in a simultaneous manner. Therefore, QTL mapping and GWAS can be technically performed using the same technology. There are two types of relatedness among individuals: population structure and cryptic relatedness. The multiple marker model (replacing polygenic control) has removed the cryptic relatedness, but the population structure remains. To incorporate population structure into the SBL model, we simply modified the fixed effect component of the model. For example, if the first two principal components are used to control the population structure, we only need to add two more columns to the design matrix of the fixed effects; no additional work is needed.

The biggest challenge for the multiple locus model is the high dimensionality of the genotype data. It is difficult to fit the number of model effects that are hundreds or even thousands times larger than n, although the large m is neither a problem for interval mapping and CIM, nor for the classical mixed-model GWAS. The limiting factor in the classical mixed-model GWAS is n because performing eigen-decomposition for a kinship matrix is daunting and even prohibited when n is very large. The FaST-LMM method (Lippert *et al.*, 2011) can handle $n = 100k$ within a reasonable time frame, it is a single marker scanning approach. The proposed SBL does not involve any large matrix calculation, and thus the sample size issue is not a threat. Our simulation studies show that the SBL program took only about 6 h to analyze a dataset with $n = 100k$ and $m = 100k$, while LASSO failed to handle such a large dataset (see Supplementary Table S2). A legitimate question is whether there is a limit of m for the SBL method to handle. The answer is – a larger n allows the model to handle a larger m. In the current GWAS populations, several millions of markers are common, but the n may only be in the order of thousands at most. For this type of data, an immediate solution is to divide the entire genome evenly into a number of LD blocks and select one marker per block for analysis. An alternative, and perhaps a better, approach is to treat each LD block as a bin and use a binned genotype to represent the LD block as suggested by Xu (2013).

A special characteristic of the classical mixed-model GWAS is the 'island' phenomenon, caused by LD, around each peak in the Manhattan plot. The multiple locus model eliminates large area of the 'island' and leaves only an isolated peak (a 'lighthouse') because each estimated effect is conditional on other effects being fitted in the model. When the effect of the peaked marker is included in the model, the effects of neighboring markers will disappear. This behavior of the multiple locus model is supposed to be a good characteristic, but it often triggers alarms in people who do not understand the difference between single locus models and multiple locus models. The simulation experiments presented in this study serve the very purpose of convincing readers to trust the isolated peaks of the multiple locus model GWAS.

Overall, the BSLMM program (Zhou *et al.*, 2013) appears to be more powerful than all methods compared (including SBL), but it is also the slowest method. For example, BSLMM took more than 9 h to complete a GWAS for $m = 5k$ markers with a sample size $n = 1495$, but SBL took only 18 s (see Supplementary Table S5). The high computational time for BSLMM may not be a problem if we only analyze just a few agronomic traits. However, QTL mapping and GWAS are being applied to thousands of metabolomic traits (Gong *et al.*, 2013; Wen *et al.*, 2014), thousands of phenomic traits (Yang *et al.*, 2014) and tens of thousands of expression traits (Wang *et al.*, 2014). It is hard to convince an investigator to run BSLMM for eQTL mapping for 20 000 transcriptomic traits while much faster programs are available.

We now discuss some theoretical basis of the SBL method. The term of SBL was first seen in Tipping (2001). He treated regression coefficients of a linear model as variables following their own distributions. When the regression coefficients are considered as parameters, each coefficient-specific normal distribution becomes a prior distribution. As a result, the method belongs to the Bayesian family. The variance of each normal prior is then estimated from the data (empirical Bayes). From a penalized regression point of view, the method implements an L_2 penalty, which is not sparse (Hoerl and Kennard, 1970; Zou and Hastie, 2005). However, Tipping (2001) used a special Gamma prior for the inverse of the variance in the normal prior to allow the estimated variance to have some mass at zero (sparseness). This technique is different from the spike-and-slab type of prior (Ishwaran and Rao, 2005; Johnstone and Silverman, 2004), which is a mixture of two distributions. Tipping (2001) method updates *one variance component* at a time conditional on *variance components* of all other effects and the model effects *per se* never occur in the model. As a result, the method is computationally demanding. The empirical Bayes method of Xu (2007) for mapping epistatic effects essentially uses an algorithm very similar to Tipping (2001).

There are three major differences between this method and Tipping's RVM. (i) Tipping's RVM is a kernel-based prediction method, where the original feature matrix ($n \times m$) is used to construct a kernel matrix ($n \times n$) (n is the sample size and m is the number of markers). Tipping's prediction model fits the kernel with a maximum of n regression coefficients but our model deals with m regression coefficients. Tipping's method is not suitable for association studies but only for prediction. Therefore, we cannot compare Tipping's RVM with our SBL. We borrowed the term 'sparse Bayesian learning' from Tipping because both are Bayesian approaches, and both can be sparse (variable selection). (ii) Tipping's RVM directly maximizes the predictability by estimating the prior variance of each regression coefficient (regression coefficients are only produced once after all variances are estimated and the iteration process converges), while our method estimates the prior variance and the regression coefficient simultaneously for each marker so that the model is very simple when estimating one regression coefficient and its variance conditional on the regression coefficients of all other markers. (iii) Tipping takes the inverse of variance as the parameter and assigns a Gamma distribution to this parameter, while we take the variance as the parameter and assign the variance an inverse Chi-squared distribution. Tipping's RVM involves two hyper parameters (a, b) and our method involves only one hyper parameter (τ).

When conditional on the posterior modes of all other effects, the model for the current variance is extremely simple because all other effects are treated as known quantities and subtracted from the observed y vector. This explains the high computational efficiency of the proposed SBL in this paper. This approach has been taken in statistics for variable selection by Johnstone and Silverman (2004) and recently by Pungpapong *et al.* (2015) who called the method an iterative conditional mode algorithm. A theoretical justification for replacing the term 'conditional on variances' by the term 'conditional on modes' can be found in Equation (4.4) of Mackay (1992).

Acknowledgements

We are grateful to two anonymous reviewers for their constructive comments and suggestions on an earlier version of the manuscript. The authors

appreciate Dr Ruidong Li from Jia's lab at UCR for his help in searching the rice genome annotation file to match the significant SNPs with known genes. The authors also thank Dr John Chater from Jia's lab at UCR for his help in proofreading the manuscript.

Funding

This project was supported by the United States National Science Foundation Collaborative Research Grant [DBI-1458515 to S.X.]; and the 'Green Super Rice for the Resource Poor Africa and Asia Phase III' Grant from the International Rice Research Institute [A-2015-50 (DRPC2015-49) to S.X.].

Conflict of Interest: none declared.

References

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Methodol.*, **57**, 289–300.

Bulik-Sullivan,B.K. *et al.* (2015) LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.*, **47**, 291–295.

Chen,J. *et al.* (2015) An evolutionarily conserved gene, FUWA, plays a role in determining panicle architecture, grain shape and grain weight in rice. *Plant J.*, **83**, 427–438.

Efron,B. and Tibshirani,R.J. (1994) *An Introduction to the Bootstrap*. CRC Press, Boca Raton.

Fan,C. *et al.* (2006) GS3, a major QTL for grain length and weight and minor QTL for grain width and thickness in rice, encodes a putative transmembrane protein. *Theor. Appl. Genet.*, **112**, 1164–1171.

Friedman,J. *et al.* (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**, 1–22.

Gong,L. *et al.* (2013) Genetic analysis of the metabolome exemplified using a rice population. *Proc. Natl. Acad. Sci. USA*, **110**, 20320–20325.

Guan,Y. and Stephens,M. (2011) Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann. Appl. Stat.*, **5**, 1780–1815.

Henderson,C.R. (1975) Best linear unbiased estimation and prediction under a selection model. *Biometrics*, **31**, 423–447.

Hoerl,A.E. and Kennard,R.W. (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Huang,X. *et al.* (2015) Genomic analysis of hybrid rice varieties reveals numerous superior alleles that contribute to heterosis. *Nat. Commun.*, **6**, 6258.

Hudson,D. *et al.* (2013) Rice cytokinin GATA transcription factor 1 regulates chloroplast development and plant architecture. *Plant Physiol.*, **162**, 132–144.

Ishwaran,H. and Rao,J.S. (2005) Spike and slab variable selection: frequentist and Bayesian strategies. *Ann. Stat.*, **33**, 730–773.

Ithnin,M. *et al.* (2017) Multiple locus genome-wide association studies for important economic traits of oil palm. *Tree Genet. Genomes*, **13**, 103.

Jansen,R.C. (1993) Interval mapping of multiple quantitative trait loci. *Genetics*, **135**, 205–211.

Jin,J. *et al.* (2015) MORF-RELATED GENE702, a reader protein of trimethylated histone H3 lysine 4 and histone H3 lysine 36, is involved in brassinosteroid-regulated growth and flowering time control in rice. *Plant Physiol.*, **168**, 1275–1285.

Johnstone,I.M. and Silverman,B.W. (2004) Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *Ann. Stat.*, **32**, 1594–1649.

Kang,H.M. *et al.* (2008) Efficient control of population structure in model organism association mapping. *Genetics*, **178**, 1709–1723.

Kao,C.-H. *et al.* (1999) Multiple interval mapping for quantitative trait loci. *Genetics*, **152**, 1203–1216.

Lin,Y. *et al.* (2016) RLS3, a protein with AAA+ domain localized in chloroplast, sustains leaf longevity in rice. *J. Integr. Plant Biol.*, **58**, 971–982.

Lippert,C. *et al.* (2011) FaST linear mixed models for genome-wide association studies. *Nat. Methods*, **8**, 833–835.

Lockhart,R. *et al.* (2014) A significance test for the lasso. *Ann. Stat.*, **42**, 413–468.

Loh,P.-R. *et al.* (2015) Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.*, **47**, 284–290.

Lv,X. *et al.* (2015) Oryza sativa chloroplast signal recognition particle 43 (OscpSRP43) is required for chloroplast development and photosynthesis. *PLoS One*, **10**, e0143249.

Ma,L. *et al.* (2017) ABNORMAL VASCULAR BUNDLES regulates cell proliferation and procambium cell establishment during aerial organ development in rice. *New Phytol.*, **213**, 275–286.

Mackay,D.J.C. (1992) Bayesian interpolation. *Neural Comput.*, **4**, 415–447.

Meuwissen,T.H.E. *et al.* (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, **157**, 1819–1829.

Ortega,J.M. and Rheinboldt,W.C. (1970) *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York.

Pungpapong,V. *et al.* (2015) Selecting massive variables using an iterated conditional modes/medians algorithm. *Electron. J. Stat.*, **9**, 1243–1266.

Purcell,S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

Segura,V. *et al.* (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.*, **44**, 825–830.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Methodol.*, **58**, 267–288.

Tipping,M.E. (2001) Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.*, **1**, 211–244.

Wang,J. *et al.* (2014) An expression quantitative trait loci-guided co-expression analysis for constructing regulatory network using a rice recombinant inbred line population. *J. Exp. Bot.*, **65**, 1069–1079.

Wen,W. *et al.* (2014) Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. *Nat. Commun.*, **5**, 3438.

Weng,J. *et al.* (2008) Isolation and initial characterization of GW5, a major QTL associated with rice grain width and weight. *Cell Res.*, **18**, 1199–1209.

Woodbury,M.A. (1950) Inverting modified matrices. *Memorandum Report 42*, p.336.

Xu,S. (2007) An empirical Bayes method for estimating epistatic effects of quantitative trait loci. *Biometrics*, **63**, 513–521.

Xu,S. (2013) Genetic mapping and genomic selection using recombination breakpoint data. *Genetics*, **195**, 1103–1115.

Xu,S. (2013) Mapping quantitative trait loci by controlling polygenic background effects. *Genetics*, **195**, 1209–1222.

Xu,S. (2017) Predicted residual error sum of squares of mixed models: an application for genomic prediction. *G3*, **7**, 895–909.

Xue,W. *et al.* (2008) Natural variation in Ghd7 is an important regulator of heading date and yield potential in rice. *Nat. Genet.*, **40**, 761–767.

Yan,D. *et al.* (2015) CURVED CHIMERIC PALEA 1 encoding an EMF 1-like protein maintains epigenetic repression of O s MADS 58 in rice palea development. *Plant J.*, **82**, 12–24.

Yang,W. *et al.* (2014) Combining high-throughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice. *Nat. Commun.*, **5**, 5087.

Yu,H. *et al.* (2011) Gains in QTL detection using an ultra-high density SNP map based on population sequencing relative to traditional RFLP/SSR markers. *PLoS One*, **6**.

Yu,J. *et al.* (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.*, **38**, 203–208.

Zang,G. *et al.* (2016) The De-Etiolated 1 homolog of arabidopsis modulates the ABA signaling pathway and ABA biosynthesis in rice. *Plant Physiol.*, **171**, 1259–1276.

Zeng,Z.-B. (1994) Precision mapping of quantitative trait loci. *Genetics*, **136**, 1457–1468.

Zhou,X. *et al.* (2013) Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genet.*, **9**, e1003264.

Zhou,X. and Stephens,M. (2012) Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.*, **44**, 821–824.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.*, **67**, 301–320.