frontiers
in Genetics

METHODS
published: 11 October 2021
doi: 10.3389/fgene.2021.717457

Single-Trait and Multiple-Trait Genomic Prediction From Multi-Class Bayesian Alphabet Models Using Biological Information

*Zigui Wang and Hao Cheng**

Department of Animal Science, University of California, Davis, Davis, CA, United States

Genomic prediction has been widely used in multiple areas and various genomic prediction methods have been developed. The majority of these methods, however, focus on statistical properties and ignore the abundant useful biological information like genome annotation or previously discovered causal variants. Therefore, to improve prediction performance, several methods have been developed to incorporate biological information into genomic prediction, mostly in single-trait analysis. A commonly used method to incorporate biological information is allocating molecular markers into different classes based on the biological information and assigning separate priors to molecular markers in different classes. It has been shown that such methods can achieve higher prediction accuracy than conventional methods in some circumstances. However, these methods mainly focus on single-trait analysis, and available priors of these methods are limited. Thus, in both single-trait and multiple-trait analysis, we propose the multi-class Bayesian Alphabet methods, in which multiple Bayesian Alphabet priors, including RR-BLUP, BayesA, BayesB, BayesCΠ, and Bayesian LASSO, can be used for markers allocated to different classes. The superior performance of the multi-class Bayesian Alphabet in genomic prediction is demonstrated using both real and simulated data. The software tool JWAS offers open-source routines to perform these analyses.

Keywords: multiple-trait, multi-class, genomic prediction, Bayesian Alphabet, biological information

OPEN ACCESS

Edited by:
Peng Xu,
Xiamen University, China

Reviewed by:
Mario Calus,
Wageningen University and Research, Netherlands
Ming Fang,
Jimei University, China

****Correspondence:***
Hao Cheng
qtlcheng@ucdavis.edu

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: *31 May 2021*
Accepted: *23 August 2021*
Published: *11 October 2021*

Citation:
Wang Z and Cheng H (2021) Single-Trait and Multiple-Trait Genomic Prediction From Multi-Class Bayesian Alphabet Models Using Biological Information. Front. Genet. 12:717457. doi: 10.3389/fgene.2021.717457

1. INTRODUCTION

Genomic prediction, proposed by Meuwissen et al. (2001), utilizes genomic information, such as single-nucleotide polymorphisms (SNPs), to estimate genotypic values or breeding values of complex traits. In the last decades, with the fast development of genotyping and sequencing technology, high-density genotype data has become much easier to access (Harris et al., 2011; Kranis et al., 2013). Accompanied by the high-density data, genomic prediction has been widely used in many areas, including animal breeding (e.g., Hayes et al., 2009a; Erbe et al., 2012), plant breeding (e.g., Wang et al., 2018; Moeinizade et al., 2020), and human disease risk prediction (e.g., Abraham et al., 2014, 2016).

A large number of genomic prediction methods with different statistical assumptions have been developed. Among these methods, genomic best linear unbiased prediction (GBLUP) (Habier et al., 2007; VanRaden, 2008; Hayes et al., 2009b), where a genomic relationship matrix is used to accommodate the covariances among breeding values, is widely used. GBLUP, however, assumes a priori that all marker effects share the same normal distribution, which may not be biologically meaningful, especially for traits controlled by a few causal variants. Furthermore, a collection of Bayesian Alphabet methods (Meuwissen et al., 2001; Fernando and Garrick, 2013; Cheng et al., 2018b; Gianola and Fernando, 2020) have been developed to incorporate different priors on marker effects, e.g., BayesA and BayesB (Meuwissen et al., 2001). Notice that GBLUP is equivalent to a Bayesian Alphabet model with a normal prior for the marker effects (Fernando, 1998; Habier et al., 2007; Strandén and Garrick, 2009). These methods, however, are still developed mainly based on statistical consideration and ignore the abundant biological information. To bridge the gap between the statistical model for genomic prediction and underlying biological architectures, researchers have proposed several methods to incorporate biological information into genomic prediction and have shown that incorporating biological information has the potential to improve the prediction accuracy in some cases (Zhang et al., 2014; Gao et al., 2015; Edwards et al., 2016).

One purpose of incorporating biological information is to relax the assumption that each locus is equally likely to affect the trait, i.e., all loci share the same prior distribution. This assumption is less biologically meaningful, e.g., some loci may be known to lead non-synonymous coding changes or have functional effects on candidate genes (MacLeod et al., 2016). One strategy to achieve this purpose is weighting markers based on the biological information and then integrating the weighting information into the model construction (Zhang et al., 2014; Gao et al., 2015). Zhang et al. (2014) incorporated the QTL list obtained in previous genome-wide association studies (GWAS) into GBLUP, i.e., when constructing genomic relationship matrix, markers were weighted based on the frequency of corresponding genomic regions being reported in the QTL list (Zhang et al., 2014). Gao et al. (2015) incorporated previous GWAS results by using locus-specific inclusion probability based on the *p*-values from GWAS.

In addition to weighting markers, another strategy to incorporate biological information is marker allocation. It has been observed that molecular markers from different genomic regions have different prediction abilities (Erbe et al., 2012; Morota et al., 2014; Do et al., 2015; Abdollahi-Arpanahi et al., 2016) and the marker allocation is beneficial if a particular class is enriched for QTL. To better fit these genomic regions with different genetic architectures, recent studies have tried to allocate genome-wide molecular markers into multiple classes based on the prior biological information and conduct genomic prediction based on these marker classes jointly. Speed and Balding (2014) proposed such a method under the GBLUP framework called MultiBLUP, which divides breeding values into multiple classes to allow different effect-size variances. A Bayesian regression method called BayesRC (MacLeod et al., 2016) was also proposed to allocate SNPs into multiple classes, where a BayesR prior was assigned to each class. It has been shown that allocating markers into different classes can improve predictive accuracy in some circumstances (Speed and Balding, 2014; MacLeod et al., 2016). The idea to allocate markers into multiple classes has also been used in a haplotype-based genomic prediction model (Xu et al., 2020), in which effects of haplotype blocks are estimated using both numerical dosage and categorical coding strategies (Martini et al., 2017) for each genomic class.

To our knowledge, most methods that allocate SNPs into different classes, focus on single-trait analysis and available priors of these methods are limited. Thus, the primary goal of this research is to present a more general Bayesian Alphabet method that can handle both single-trait and multiple-trait analysis, while is able to assign multiple Bayesian Alphabet priors, including RR-BLUP, BayesA, BayesB, BayesCΠ, and Bayesian LASSO, to markers in different SNP classes. The new genomic prediction method we implemented is called multi-class Bayesian Alphabet, where the term "Bayesian Alphabet" denotes a collection of Bayesian Alphabet priors adopted for marker effects. Our multi-class Bayesian Alphabet works for both single-trait and multiple-trait analysis. The performance of the multi-class Bayesian Alphabet is studied using real and simulated data.

2. MATERIALS AND METHODS

2.1. Multi-Class Bayesian Alphabet Models

For simplicity, the general mean is assumed as the only fixed effect, thus the general form of the multi-class Bayesian Alphabet model for *i*th genotyped observation can be written as:

$$\boldsymbol{y}_i = \boldsymbol{\mu} + \sum_{l=1}^{g} \sum_{f_l \in C_l} m_{if_l} \boldsymbol{\alpha}_{f_l} + \boldsymbol{e}_i \tag{1}$$

where $\boldsymbol{y}_i$ is a vector of phenotypic values of t traits for observation i; $\boldsymbol{\mu}$ is a vector of overall means for t traits; m_{if_l} is the genotype covariate at locus f_l (coded as 0,1,2) in SNP class C_l for observation i; g is the number of SNP classes; $\boldsymbol{\alpha}_{f_l}$ is a vector of the corresponding allele substitution effects (marker effects) of t traits for locus f_l; and $\boldsymbol{e}_i$ is a vector of residuals for observation i. Note that when the number of traits $t = 1$, the general form above simplifies to the single-trait model, and all vectors of effects in Equation 1 become scalars. The fixed effect $\boldsymbol{\mu}$ is assigned a flat prior. The residuals, $\boldsymbol{e}_i$, are a priori assumed to be independently and identically distributed multivariate normal vectors with null mean and covariance matrix $\boldsymbol{R}$, which is assigned an inverse Wishart prior distribution, $\boldsymbol{W}^{-1}(\boldsymbol{S}_e, \nu_e)$, with degrees of freedom $\nu_e = 4$ and scale matrix $\boldsymbol{S}_e$ such that the prior mean of $\boldsymbol{R}$ equals half of the phenotypic variance. Note that when number of traits $t = 1$, the prior for $\boldsymbol{R}$ follows a scaled inverted chi-square distribution.

To incorporate known biological information, marker effects of SNPs in the same class are assumed to have identical Bayesian Alphabet prior. Different from conventional Bayesian Alphabet methods, our multi-class Bayesian Alphabet methods allow assigning different Bayesian Alphabet priors to marker effect $\boldsymbol{\alpha}_{f_l}$ in different SNP classes. These priors are discussed in the following section 2.2.

2.2. Bayesian Prior for Marker Effects

Multiple priors are implemented in our multi-class Bayesian Alphabet models, including BayesA, BayesB, BayesCΠ, RR-BLUP, and Bayesian LASSO. In multiple-trait analysis, with BayesB and BayesCΠ priors, each locus is allowed to affect any combination of traits (Cheng et al., 2018b). In multiple-trait BayesB and BayesCΠ, the vector of marker effects at locus f_l can be written as $\boldsymbol{\alpha}_{f_l} = \boldsymbol{D}_{f_l}\boldsymbol{\beta}_{f_l}$, where $\boldsymbol{D}_{f_l}$ is a diagonal matrix whose diagonal elements are $\boldsymbol{\delta}_{f_l} = (\delta_{f_l1}, \delta_{f_l2}..., \delta_{f_lt})$, where δ_{f_lt} is an indicator variable indicating whether the marker effect of locus f_l for trait t is zero or not. We use numeric labels "1," "2,"$\cdots$, "z" to represent all possible combinations for $\boldsymbol{\delta}_{f_l}$, in which case the prior distribution for $\boldsymbol{\delta}_{f_l}$ is: $p(\boldsymbol{\delta}_{f_l}$ = "i") = $\Pi_1 I(\boldsymbol{\delta}_{f_l} =$ "1") + $\Pi_2 I(\boldsymbol{\delta}_{f_l}$ = "2") +...+ $\Pi_l I(\boldsymbol{\delta}_{f_l} =$ "z") where Π_i is the prior probability that the vector $\boldsymbol{\delta}_{f_l}$ corresponds to the vector labeled "i" and $\sum \Pi_i = 1$. A uniform prior distribution is assigned to $\Pi = (\Pi_1, \Pi_2, ...\Pi_l)$ (Cheng et al., 2018b). In multiple-trait BayesB, the prior for $\boldsymbol{\beta}_{f_l}$ is a multivariate normal distribution with null mean and locus-specific covariance matrix $\boldsymbol{G}_{f_l}$, which is assigned an inverse Wishart prior, $W_t^{-1}(\boldsymbol{S}_\beta, \nu_\beta)$. In multiple-trait BayesCΠ, instead of locus-specific covariance matrix $\boldsymbol{G}_{f_l}$, $\boldsymbol{\beta}_{f_l}$ is assumed to follow a multivariate normal prior with null mean and common covariance matrix $\boldsymbol{G}$, which is assumed to have an inverse Wishart prior distribution, $W^{-1}(\boldsymbol{S}_\beta, \nu_\beta)$, with degrees of freedom $\nu_\beta = 4$ and scale matrix $\boldsymbol{S}_\beta$ such that the prior mean of genetic variance equals half of the phenotypic variance. In single-trait analysis, $\boldsymbol{D}_{f_l}$, $\boldsymbol{G}_{f_l}$, and marker effect $\boldsymbol{\beta}_{f_l}$ become scalars. The prior of β_{f_l} becomes a univariate normal distribution; the prior of G_{f_l} becomes an inverted chi-square distribution, and D_{f_l} is an indicator variable indicating whether the maker effect is zero or not. In both single-trait and multiple-trait analysis, BayesA and RR-BLUP are just special cases of BayesB and BayesCΠ respectively, where all markers are assumed to have effects on all traits (Fernando and Garrick, 2013). The Bayesian LASSO prior is also included in the multi-class Bayesian Alphabet. In Bayesian LASSO, the multivariate Laplace prior distribution with a null mean is assigned to marker effect vector $\boldsymbol{\alpha}_{f_l}$ (Gianola and Fernando, 2020) in multiple-trait analysis. In single-trait Bayesian LASSO, the prior for α_{f_l} is a double exponential distribution (Tibshirani, 1996; Gianola, 2013).

2.3. Data Analysis

2.3.1. Real Data

Two public datasets are used to evaluate the performance of multi-class Bayesian Alphabet models. The first dataset, which is used to evaluate the single-trait analysis, is composed of genotypic and phenotypic data from Michigan State University Pig Resource Population (MSUPRP) raised at the Michigan State University Swine Teaching and Research Farm, East Lansing,MI (Edwards et al., 2008). After quality control (Duarte et al., 2014), 928 individuals and 42,246 SNPs remain. The trait *13-week tenth rib backfat (mm)* is considered in this analysis. The original data is available at https://msu.edu/~steibelj/JP_files/GBLUP.html. The genome annotation information for the pig dataset used in this paper is obtained from the Ensembl (Rainer et al., 2019) database using the GALLO package (Fonseca et al., 2020). Five annotation regions are identified in the pig dataset, and will be used in our analysis. The number of SNPs in the protein coding, RNA, processed pseudogene, intergenic, and pseudogene regions are 15084, 1840, 107, 24838, and 377, respectively.

The second dataset, which is used to evaluate the multiple-trait analysis, is from the Rice Diversity Panel with 370 *Oryza sativa* individual accessions (Zhao et al., 2011). Three traits *plant height (PH)*, *flowering time in Arkansas (FTA)*, and *panicle number per plant (PN)* are considered. After removing the genotypes missing for these three traits or with minor allele frequency < 0.05, 33,519 SNPs are included in our analysis. The phenotypic and genotypic data are publicly available at http://www.ricediversity.org/. The genome annotation information for the rice dataset is obtained from Ensembl (Rainer et al., 2019) database using the biomart package (Durinck et al., 2009). Four annotation regions are identified in the rice dataset, and will be used in our analysis. The number of SNPs in protein coding, RNA, non-translating CDS, and intergenic regions are 14129, 3, 176, and 19211, respectively.

We identified total 6 genomic annotations: protein coding, processed pseudogene, pseudogene, non-coding RNA, non-translating CDS, and intergenic. According to Howe et al. (2020), the "protein coding" class is comprised of the SNPs within the gene that contains an open reading frame (ORF). In other words, these SNPs may be processed into messenger RNAs (mRNAs) which, after their export to the cytosol, are translated into proteins (Harrow et al., 2009). The "pseudogene" class contains SNPs within the genes that have coding-sequence deficiencies like frameshifts and premature stop codons but resemble protein-coding genes (Howe et al., 2020; Tutar, 2012). The "processed pseudogene" class includes the SNPs in the pseudogene that lack introns and is thought to arise from reverse transcription of messenger RNA followed by reinsertion of DNA into the genome (Howe et al., 2020). The "non-coding RNA" class contains SNPs within RNA that are not translated into a protein (Howe et al., 2020). The "non-translating CDS" class represents SNPs in coding sequence regions that are not translated to a protein (Howe et al., 2020). All other SNPs were allocated to the class "intergenic".

2.3.2. Simulated Data

To comprehensively compare multi-class Bayesian Alphabet with conventional Bayesian Alphabet for genomic prediction, we conducted simulations based on the real genotypes from Michigan State University Pig Resource Population (MSUPRP) described above (Edwards et al., 2008). The simulation strategies in MacLeod et al. (2016) were applied. 500 QTLs were randomly selected from SNP class "protein coding", i.e., SNPs with the annotation "protein coding". In addition, 20 QTLs were randomly selected across the genome. The same QTL positions were used in our simulation. Two correlated traits of heritabilities equal to 0.5 and 0.9 were simulated, where pleiotropic QTL effects were sampled from a multivariate normal distribution with null mean and covariance matrix $\boldsymbol{G} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$. The trait of heritability 0.5 was used in our single-trait analysis, and both traits were used in our multiple-trait analysis. There were total

TABLE 1 | Mean prediction accuracy comparison between conventional and multi-class Bayesian Alphabet on single-trait simulated data.

Method	RR-BLUP	BayesA	BayesB	BayesCπ	Bayesian LASSO	Ensemble
Conventional	0.542*	0.542	0.547*	0.547*	0.541*	0.545*
Multi-class	0.563*	0.542	0.565*	0.565*	0.563*	0.563*

*The comparison of mean prediction accuracy across 150 single-trait validation datasets (30 simulated data × five-fold cross validation) between multi-class Bayesian Alphabet using genome annotation information and conventional Bayesian Alphabet. The paired t-test ($p < 0.1$) was used to declare the significant difference. *Denotes that significant differences were found between multi-class Bayesian Alphabet and conventional Bayesian Alphabet with RR-BLUP, BayesB, BayesCπ Bayesian LASSO, and ensemble approach, respectively ($p < 0.1$).*

TABLE 2 | Mean prediction accuracy comparison between conventional and multi-class Bayesian Alphabet on multiple-trait simulated data.

Method	RR-BLUP	BayesA	BayesB	BayesCΠ	Bayesian LASSO	Ensemble
Conventional	0.552*	0.554	0.565*	0.564*	0.552*	0.561*
Multi-class	0.572*	0.553	0.578*	0.577*	0.572*	0.575*

*The comparison of mean prediction accuracy across 150 multiple-trait validation datasets (30 simulated data × five-fold cross validation) between multi-class Bayesian Alphabet using genome annotation information and conventional Bayesian Alphabet. The paired t-test ($p < 0.1$) was used to declare the significant difference. *Denotes that significant differences were found between multi-class Baysian Alphabet and conventional Bayesian Alphabet with RR-BLUP, BayesB, BayesCπ, Baysian LASSO, and ensemble approach, respectively ($p < 0.1$).*

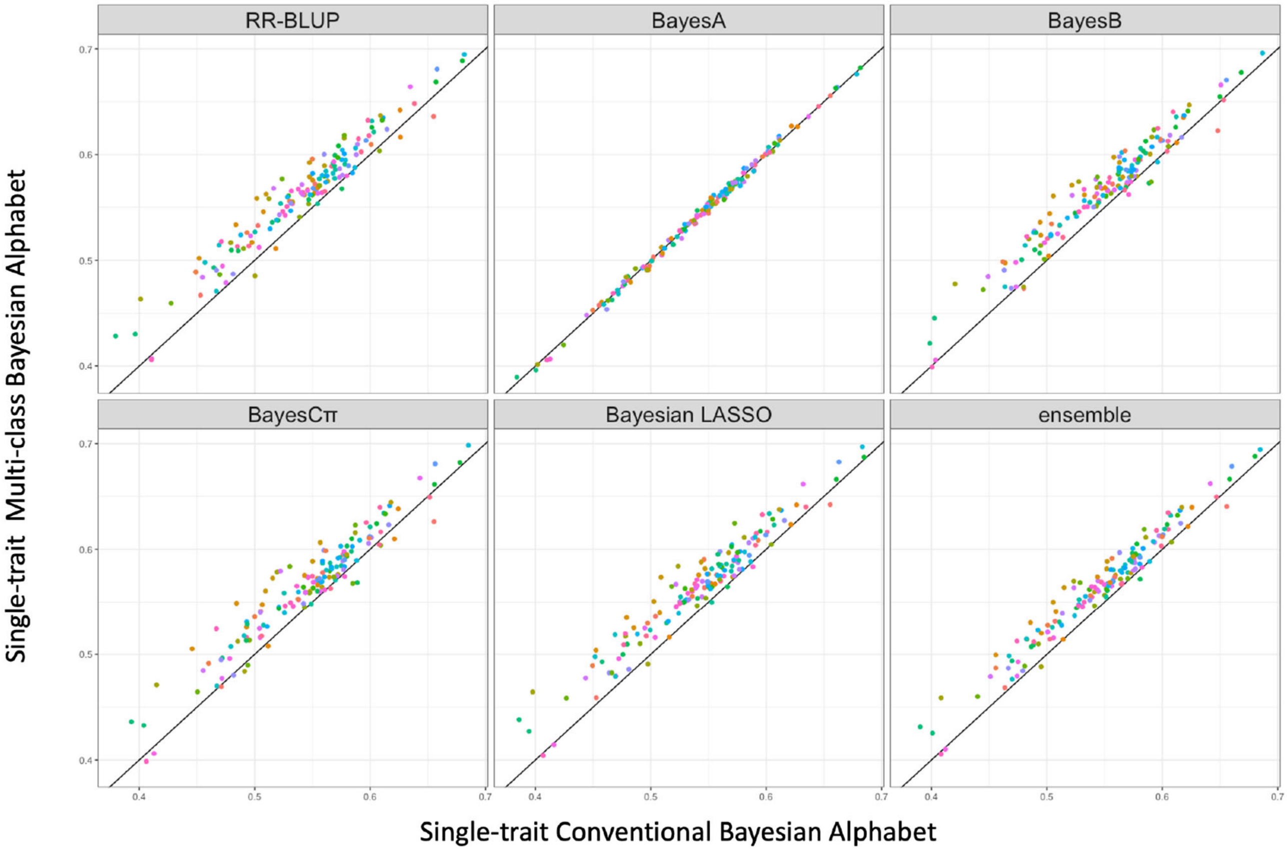

FIGURE 1 | The pairwise predication accuracy comparison between conventional and multi-class Bayesian Alphabet on single-trait simulated data. The 30 simulated datasets were distinguished by color; the x-axis represents the genomic prediction accuracy obtained from conventional Bayesian Alphabet methods; the y-axis represents the genomic prediction accuracy obtained from multi-class Bayesian Alphabet methods; the diagonal line is used for reference such that a dot above the line represents a validation with higher accuracy for multi-class Bayesian Alphabet. Significant differences were found between multi-class Bayesian Alphabet and conventional Bayesian Alphabet with RR-BLUP, BayesB, BayesCπ Bayesian LASSO, and ensemble approach, respectively ($p < 0.1$).

30 different datasets being simulated based on the simulation processes described above.

2.3.3. Cross Validation

The dataset was randomly split into training and validation datasets following an 8:2 ratio for each replicate. 50 replicates and 5 replicates were applied to the real and simulated datasets, respectively. The prediction accuracy was calculated as the mean Pearson correlation between the estimated breeding values and phenotypic records of observations in validation datasets. Conventional and multi-class Bayesian Alphabet methods were compared using RR-BLUP, BayesA, BayesB, BayesCΠ, and Bayesian LASSO priors. In addition to the above five Bayesian methods, an ensemble approach that uses average estimated breeding values across five Bayesian methods, was used to integrate multiple predictions into one summary prediction (Azodi et al., 2019).

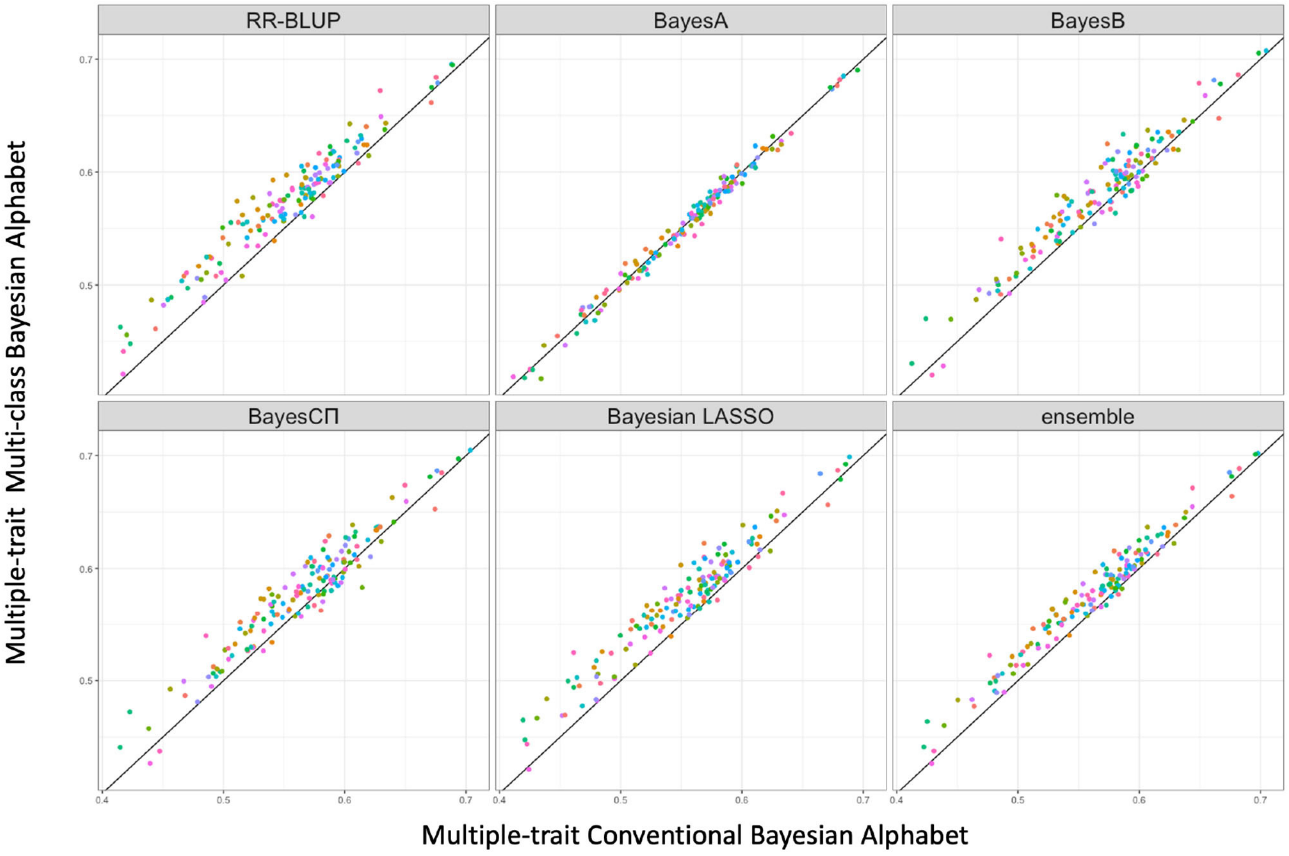

FIGURE 2 | The pairwise predication accuracy comparison between conventional Bayesian Alphabet and multi-class Bayesian Alphabet on multiple-trait simulated data. The 30 simulated datasets were distinguished by color; the x-axis represents the genomic prediction accuracy obtained from conventional Bayesian Alphabet method; the y-axis represents the genomic prediction accuracy obtained from multi-class Bayesian Alphabet method; the diagonal line is used for reference such that a dot above the line represents a validation with higher accuracy for multi-class Bayesian Alphabet. Significant differences were found between multi-class Bayesian Alphabet and conventional Bayesian Alphabet with RR-BLUP, BayesB, BayesCπ Bayesian LASSO, and ensemble approach, respectively ($p < 0.1$).

TABLE 3 | Mean prediction accuracy comparison between conventional and multi-class Bayesian Alphabet for real pig data (single-trait) and real rice data (multiple-trait).

Data	Method	RR-BLUP	BayesA	BayesB	BayesCΠ	Bayesian LASSO	Ensemble
Pig	Conventional	0.516	0.565	0.568	0.532	0.517	0.550
	Multi-class	0.516	0.565	0.569	0.532	0.516	0.550
Rice	Conventional	0.378	0.353	0.372	0.384	0.378	0.377
	Multi-class	0.374	0.357	0.363	0.375	0.373	0.373

The comparison of mean prediction accuracy on the trait 13-week tenth rib backfat (mm) from pig data and trait FTA of rice real data across 50 validation datasets between multi-class Bayesian Alphabet using genome annotation information and conventional Bayesian Alphabet. The paired t-test ($p < 0.1$) was used to declare the significant difference. No significant differences were found between multi-class and conventional Bayesian Alphabet methods for both real pig and rice data ($p < 0.1$).

The molecular markers were allocated into multiple classes using the genome annotation information. SNP classes were defined using the genome annotation information, i.e., SNPs with the same genome annotation were allocated in one class.

We have implemented these methods in JWAS (Cheng et al., 2018a), an open-source package for single-trait and multiple-trait genome-enabled prediction and analyses. The software tool JWAS offers open-source routines to perform these analyses. The documentation and examples of JWAS can be found at https://github.com/reworkhow/JWAS.jl. MCMC chains of length 100,000 with a burn-in of the first 50,000 iterations were used. The Gelman-Rubin test (Gelman and Rubin, 1992) has been used to verify the convergence of the MCMC chain.

3. RESULT

3.1. Simulated Data

Multi-class Bayesian Alphabet methods using genome annotation information were performed for both single-trait and multiple-trait prediction on the simulated data. In both single-trait and multiple-trait analysis, 5-fold cross validation was applied on 30 simulated datasets. The comparisons between multi-class and conventional Bayesian Alphabet methods are shown in **Table 1** for single-trait analysis and **Table 2** for multiple-trait analysis. The pairwise comparisons across all 30 simulated datasets are also shown in **Figure 1** for single-trait analysis and **Figure 2** for multiple-trait analysis. The 30 simulated datasets are distinguished by color. The paired t-test with a significance level 0.1 is used to declare the significant difference between prediction accuracies from multi-class and conventional Bayesian Alphabet methods.

In the single-trait analysis, significant differences in prediction accuracies were detected between multi-class and conventional Bayesian Alphabet methods with RR-BLUP, BayesB, BayesCπ, Bayesian LASSO priors, and the ensemble approach ($p < 0.1$). In detail, the mean prediction accuracies of multi-class Bayesian Alphabet were higher than conventional Bayesian Alphabet in 30 out of all 30 datasets with RR-BLUP, BayesB, BayesCπ, Bayesian LASSO priors, and ensemble approach. Multi-class Bayesian Alphabet significantly outperforms conventional Bayesian Alphabet in the ensemble approach due to the better performance of multi-class Bayesian Alphabet using these 4 priors.

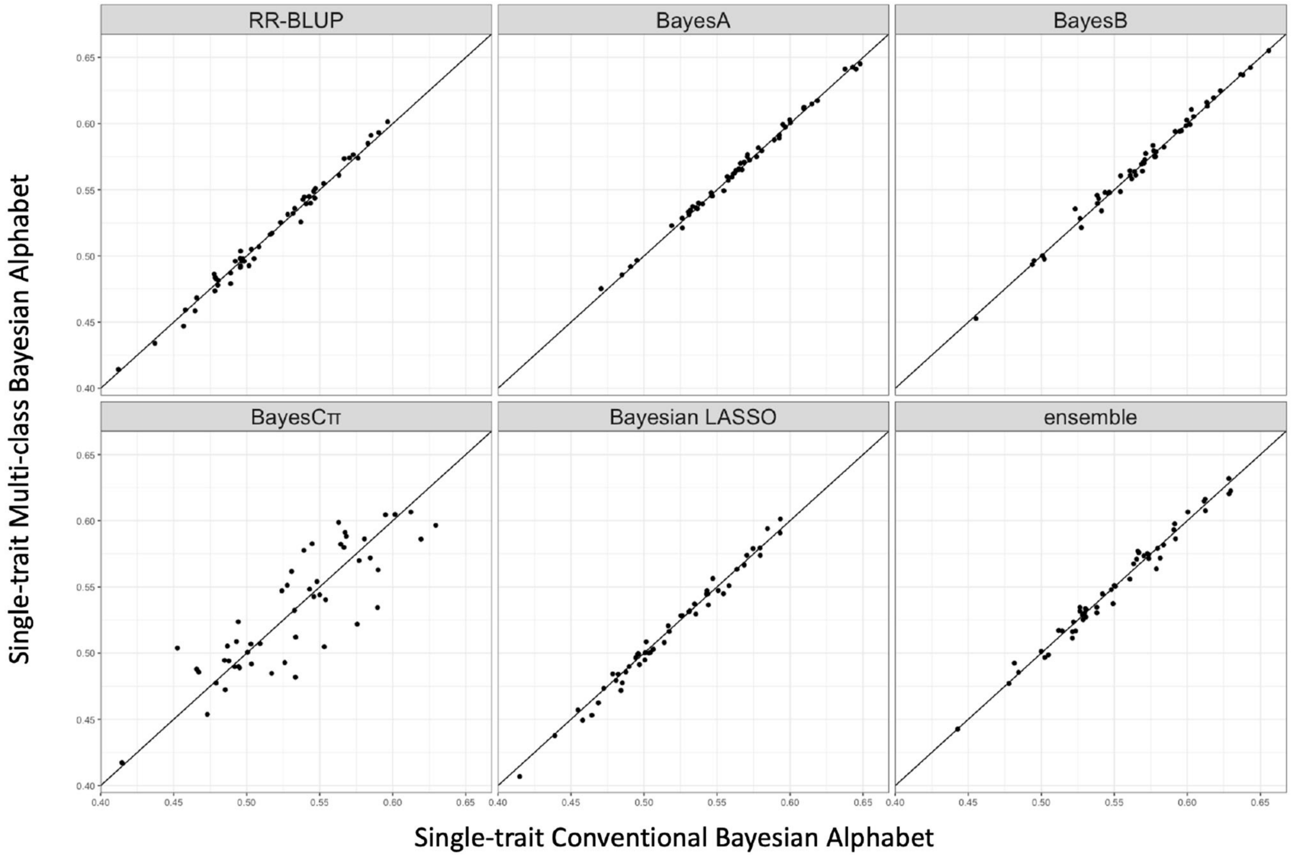

FIGURE 3 | The pairwise prediction accuracy comparisons between conventional Bayesian method and multi-class Bayesian Alphabet using genome annotation classes for the trait *13-week tenth rib backfat (mm)* from pig real data. The x-axis represents the genomic prediction accuracy obtained from conventional Bayesian method; the y-axis represents the genomic prediction accuracy obtained from multi-class Bayesian Alphabet method; the diagonal line is used for reference such that a dot above the line represents a validation with higher accuracy for multi-class Bayesian Alphabet. No significant differences were found ($p < 0.1$).

In the multiple-trait analysis, no significant differences were observed for the higher heritability trait, and results for the lower heritability trait were presented. Overall, higher prediction accuracies were usually observed for the same prior in multiple-trait analysis compared to single-trait analysis. A significant difference in prediction accuracies was detected between multi-class and conventional Bayesian Alphabet methods with RR-BLUP, BayesB, BayesCΠ, Bayesian LASSO prior ($p < 0.1$) as well as the ensemble approach. Similar to single-trait simulation result, the mean prediction accuracies of multi-class Bayesian Alphabet were higher than conventional Bayesian Alphabet in 30 out of all 30 simulated datasets with RR-BLUP, BayesB, BayesCπ Bayesian LASSO priors and the ensemble approach. The simulated data result shows that the multi-class Bayesian Alphabet has the potential to improve the prediction accuracy for both single-trait and multiple-trait analysis.

3.2. Real Data

Multi-class Bayesian Alphabet methods were performed on the pig data (Edwards et al., 2008) for single-trait analysis and the rice data (Zhao et al., 2011) for multiple-trait analysis. In the multiple-trait analysis, three traits *PH*, *FTA* and *PN* showed similar patterns on the comparison between conventional and multi-class Bayesian Alphabet methods, so only results of trait *FTA* were presented for simplicity. In both single-trait and multiple-trait analysis, 50-fold cross validation was applied. The comparison between multi-class and conventional Bayesian Alphabet methods are shown in **Table 3** for single-trait analysis and multiple-trait analysis. The pairwise comparisons across all 50 validation datasets are also shown in **Figure 3** for single-trait analysis, and **Figure 4** for multiple-trait analysis. The paired t-test with a significance level 0.1 was used to declare the significant difference between prediction accuracies from multi-class and conventional Bayesian Alphabet methods.

As shown in **Table 3**, in both real pig (single-trait) and rice (multiple-trait) data analysis, the prediction accuracies of multi-class Bayesian Alphabet using genome annotation information were not significantly different from conventional Bayesian Alphabet methods for all priors and ensemble approach.

We further studied the effect of SNP allocation on prediction accuracy by using other types of known biological information. For example, we allocated SNPs on the same chromosome to the same class such that number of chromosomes classes are fitted in multi-class Bayesian Alphabet methods. As shown

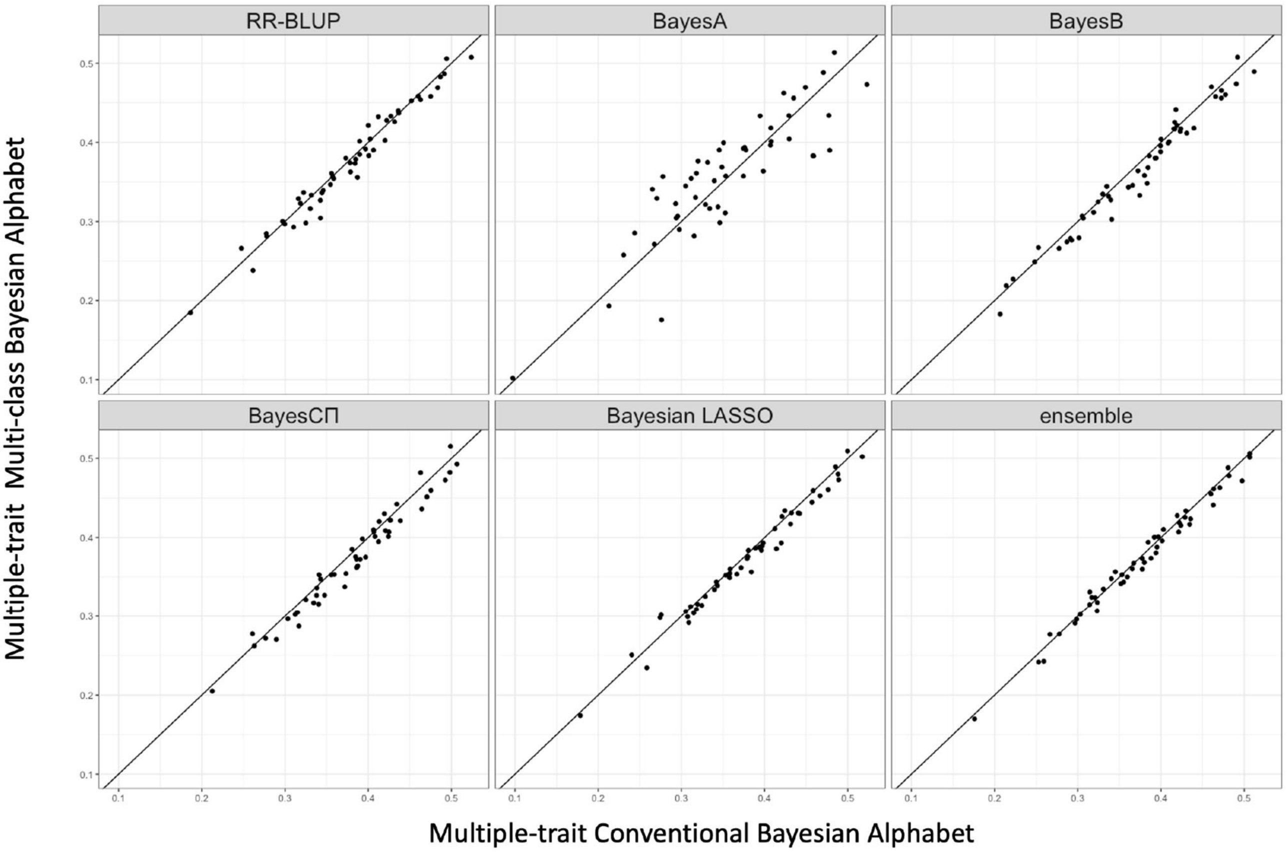

FIGURE 4 | The pairwise accuracy comparisons between conventional Bayesian method and multi-class Bayesian Alphabet using genome annotation classes for the trait *FTA* of rice real data. The x-axis represents the genomic prediction accuracy obtained from conventional Bayesian method; the y-axis represents the genomic prediction accuracy obtained from multi-class Bayesian Alphabet method; the diagonal line is used for reference such that a dot above the line represents a validation with higher accuracy for multi-class Bayesian Alphabet. No significant differences were found ($p < 0.1$).

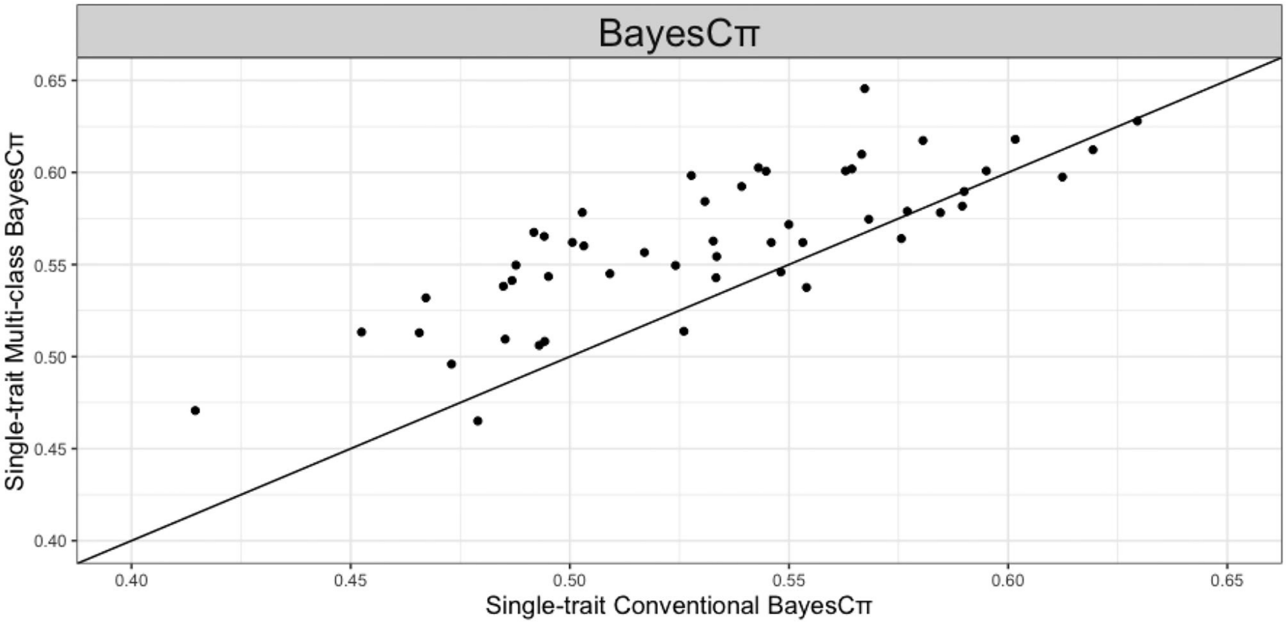

FIGURE 5 | The pairwise prediction accuracy comparisons of conventional BayesCπ and multi-class BayesCπ using chromosome classes for the trait *13-week tenth rib backfat (mm)* of real pig data. The x-axis represents the genomic prediction accuracy obtained from conventional BayesCπ; the y-axis represents the genomic prediction accuracy obtained from multi-class BayesCπ; the diagonal line is used for reference such that a dot above the line represents a validation with higher accuracy for multi-class BayesCπ. Multi-class BayesCπ significantly outperformed the conventional BayesCπ ($p < 0.1$).

in **Figure 5**, in the real pig (single-trait) data analysis, when BayesCπ prior is used, multi-class Bayesian Alphabet using chromosome classes has significantly higher prediction accuracy than the conventional Bayesian Alphabet ($p < 0.1$). To further understand why higher prediction accuracy is achieved in multi-class BayesCπ using chromosome classes, a genome-wide association study (GWAS) was performed on the same dataset, and one significant signal was detected on chromosome 6 (Chen et al., 2017). Thus, we ran another multi-class Bayesian alphabet analysis by allocating SNPs on chromosome 6 to one class and the remaining to another for a 2-class Bayesian Alphabet analysis. Higher prediction accuracy was observed in this 2-class Bayesian Alphabet analysis. It indicates that assigning SNPs into classes based on GWAS results may be one useful strategy to incorporate biological information.

4. DISCUSSION

Most genomic prediction methods usually assume all marker effects share the same prior distribution. This assumption, however, is not biologically meaningful and may potentially reduce the prediction performance when genetic architectures vary across different genomic regions (Speed and Balding, 2014). To address this issue, some methods such as MultiBLUP (Speed and Balding, 2014) and BayesRC (MacLeod et al., 2016) were proposed to allocate markers into different classes, and the superior performances of these methods were observed. Most of these methods, however, focus on single-trait analysis and have limitations in the priors used for marker effects. Thus, in this study, we presented the multi-class Bayesian Alphabet methods, which can perform both single-trait and multiple-trait analysis and provide multiple Bayesian Alphabet priors for markers allocated to different classes.

The effect of allocating markers into different classes on genomic prediction has been studied in some previous studies (Morota et al., 2014; Speed and Balding, 2014; MacLeod et al., 2016; Xu et al., 2020). Different effect-size prior distributions are assigned to molecular markers being split into multiple classes based on genetic architectures. In this paper, we use genome annotation to allocate markers into multiple classes. Note that given the different biological information, the number of classes and markers inside each class might be different. For example, we can use the GWAS results, like Zhang et al. (2014) and Gao et al. (2015), to allocate markers into two classes: one with identified causal variants and another class with the remaining markers.

The comparisons between prediction accuracies from multi-class and conventional Bayesian Alphabet are shown in **Tables 1–3**. Multi-class Bayesian Alphabet performs consistently equivalent to or better than conventional Bayesian Alphabet in both real and simulated datasets. The different performances of the multi-class Bayesian Alphabet may be caused by the genetic architectures across different genomic regions in the datasets. The methods that allocate markers into different classes outperform the conventional methods because these methods allow different priors on marker effects according to genetic architectures (Speed and Balding, 2014; MacLeod et al., 2016). If genetic architectures are similar across the SNP classes, assigning different priors will not bring significant improvement. For example, in comparisons without much difference between multi-class and conventional methods, e.g., multi-class BayesCπ using genome annotation information in the real pig data analysis, relatively small range (0.0001 to 0.03) for the estimated marker effect variances was observed across SNP classes. However, in comparison with significant differences, e.g., multi-class BayesCπ using chromosome information in the real pig data analysis, relatively large range (0.0001 to 0.15) for the estimated marker effect variances was observed across SNP classes.

Our multi-class Bayesian Alphabet method allows the coexist of the different types of priors in one model. For example, a BayesA prior can be assigned to one SNP class and a BayesCΠ prior to another. In addition, the same marker can be allocated to multiple SNP classes. Compared to other methods that allocate markers into multiple classes, our multi-class Bayesian Alphabet provides more flexibility for model construction given the genetic architectures of the traits of interest and increasing biological knowledge on the genome for both single-trait and multiple-trait analysis. However, a naive comparison among multiple multi-class Bayesian Alphabet methods is computationally intensive. For example, with 6 SNP classes and 5 types of prior, there are 5^6 possible combinations, and the computational intensity increases dramatically as the number of SNP classes grows. An efficient algorithm to choose biologically meaningful priors for each SNP class, is needed. In addition, biological knowledge generated from other projects may help to narrow down the prior candidates for each SNP class. In our multi-class Bayesian Alphabet methods tested in this paper, where computational intensities are similar to conventional methods, equivalent or better performances are consistently observed. Given that our single-trait and multiple-trait multi-class Bayesian Alphabet methods are biologically meaningful and their implementation is available in an open-source package, we expect it would be widely adopted for genomic prediction.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

HC conceived the study. HC and ZW implemented the method. ZW undertook the analysis and wrote the draft. Both authors contributed to the final version of the manuscript, read, and approved the final manuscript.

FUNDING

This work was supported by the United States Department of Agriculture, Agriculture and Food Research Initiative National Institute of Food and Agriculture Competitive Grant Nos. 2018-67015-27957 and 2021-67015-33412.

ACKNOWLEDGMENTS

We want to thank Tianjing Zhao for her suggestions on the SNP allocation and Jiayi Qu for her advice on the manuscript writing.

REFERENCES

Abdollahi-Arpanahi, R., Morota, G., Valente, B. D., Kranis, A., Rosa, G. J. M., and Gianola, D. (2016). Differential contribution of genomic regions to marked genetic variation and prediction of quantitative traits in broiler chickens. *Genet. Select. Evol.* 48:10. doi: 10.1186/s12711-016-0187-z

Abraham, G., Havulinna, A. S., Bhalala, O. G., Byars, S. G., Livera, A. M. D., Yetukuri, L., et al. (2016). Genomic prediction of coronary heart disease. *Eur. Heart J.* 37, 3267–3278. doi: 10.1093/eurheartj/ehw450

Abraham, G., Tye-Din, J. A., Bhalala, O. G., Kowalczyk, A., Zobel, J., and Inouye, M. (2014). Accurate and robust genomic prediction of celiac disease using statistical learning. *PLoS Genet.* 10:e1004137. doi: 10.1371/journal.pgen.1004137

Azodi, C. B., Bolger, E., McCarren, A., Roantree, M., de los Campos, G., and Shiu, S.-H. (2019). Benchmarking parametric and machine learning models for genomic prediction of complex traits. *G3 (Bethesda).* 9, 3691–3702. doi: 10.1534/g3.119.400498

Chen, C., Steibel, J. P., and Tempelman, R. J. (2017). Genome-wide association analyses based on broadly different specifications for prior distributions, genomic windows, and estimation methods. *Genetics* 206, 1791–1806. doi: 10.1534/genetics.117.202259

Cheng, H., Fernando, R., and Garrick, D. (2018a). "Jwas: Julia implementation of whole-genome analyses software," in *Proceedings of the World Congress on Genetics Applied to Livestock Production* (Auckland).

Cheng, H., Kizilkaya, K., Zeng, J., Garrick, D., and Fernando, R. (2018b). Genomic prediction from multiple-trait bayesian regression methods using mixture priors. *Genetics* 209, 89–103. doi: 10.1534/genetics.118.300650

Do, D. N., Janss, L. L. G., Jensen, J., and Kadarmideen, H. N. (2015). SNP annotation-based whole genomic prediction and selection: an application to feed efficiency and its component traits in pigs. *J. Anim. Sci.* 93, 2056–2063. doi: 10.2527/jas.2014-8640

Duarte, J. L. G., Cantet, R. J., Bates, R. O., Ernst, C. W., Raney, N. E., and Steibel, J. P. (2014). Rapid screening for phenotype-genotype associations by linear transformations of genomic evaluations. *BMC Bioinformatics* 15:246. doi: 10.1186/1471-2105-15-246

Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/bioconductor package biomart. *Nat. Protoc.* 4, 1184–1191. doi: 10.1038/nprot.2009.97

Edwards, D. B., Ernst, C. W., Raney, N. E., Doumit, M. E., Hoge, M. D., and Bates, R. O. (2008). Quantitative trait locus mapping in an F2 Duroc x Pietrain resource population: II. Carcass and meat quality traits. *J. Anim. Sci.* 86, 254–266. doi: 10.2527/jas.2006-626

Edwards, S. M., Sarensen, I. F., Sarup, P., Mackay, T. F. C., and Sarensen, P. (2016). Genomic prediction for quantitative traits is improved by mapping variants to gene ontology categories in *Drosophila melanogaster*. *Genetics* 203, 1871–1883. doi: 10.1534/genetics.116.187161

Erbe, M., Hayes, B., Matukumalli, L., Goswami, S., Bowman, P., Reich, C., et al. (2012). Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J. Dairy Sci.* 95, 4114–4129. doi: 10.3168/jds.2011-5019

Fernando, R. (1998). "Genetic evaluation and selection using genotypic, phenotypic and pedigree information," in *6th Wld. Cong. Genet. App. Liv. Prod.* (Armidale, NSW), 329–336.

Fernando, R. L., and Garrick, D. (2013). Genome-wide association studies and genomic prediction. *Methods Mol. Biol.* 1019, 237–274. doi: 10.1007/978-1-62703-447-0_10

Fonseca, P., Suarez-Vega, A., Marras, G., and Canovas, A. (2020). GALLO: An R package for genomic annotation and integration of multiple data sources in livestock for positional candidate loci. *GigaScience* 9:giaa149.

Gao, N., Li, J., He, J., Xiao, G., Luo, Y., Zhang, H., et al. (2015). Improving accuracy of genomic prediction by genetic architecture based priors in a Bayesian model. *BMC Genetics* 16:120. doi: 10.1186/s12863-015-0278-9

Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Stat. Sci.* 7, 457–472. doi: 10.1214/ss/1177011136

Gianola, D. (2013). Priors in whole-genome regression: the Bayesian alphabet returns. *Genetics* 194, 573–596. doi: 10.1534/genetics.113.151753

Gianola, D., and Fernando, R. L. (2020). A multiple-trait bayesian lasso for genome-enabled analysis and prediction of complex traits. *Genetics* 214, 305–331. doi: 10.1534/genetics.119.302934

Habier, D., Fernando, R. L., and Dekkers, J. C. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177, 2389–2397. doi: 10.1534/genetics.107.081190

Harris, B., Creagh, F., Winkelman, A., and Johnson, D. (2011) Experiences with the illumina high density bovine beadchip. *Interbull Bulletin*.

Harrow, J., Nagy, A., Reymond, A., Alioto, T., Patthy, L., Antonarakis, S. E., et al. (2009). Identifying protein-coding genes in genomic sequences. *Genome Biol.* 10:201. doi: 10.1186/gb-2009-10-1-201

Hayes, B., Bowman, P., Chamberlain, A., and Goddard, M. (2009a). Invited review: genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci.* 92, 433–443. doi: 10.3168/jds.2008-1646

Hayes, B. J., Visscher, P. M., and Goddard, M. E. (2009b). Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res.* 91, 47–60. doi: 10.1017/S0016672308009981

Howe, K. L., Achuthan, P., Allen, J., Allen, J., and Alvarez-Jarreta, J. (2020). Ensembl 2021. *Nucleic Acids Res.* 49, D884–D891.

Kranis, A., Gheyas, A. A., Boschiero, C., Turner, F., Yu, L., Smith, S., et al. (2013). Development of a high density 600K SNP genotyping array for chicken. *BMC Genomics* 14:59. doi: 10.1186/1471-2164-14-59

MacLeod, I. M., Bowman, P. J., Jagt, C. J. V., Haile-Mariam, M., Kemper, K. E., Chamberlain, A. J., et al. (2016). Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC Genomics* 17:144. doi: 10.1186/s12864-016-2443-6

Martini, J. W. R., Gao, N., Cardoso, D. F., Wimmer, V., Erbe, M., Cantet, R. J. C., et al. (2017). Genomic prediction with epistasis models: on the marker-coding-dependent performance of the extended GBLUP and properties of the categorical epistasis model (CE). *BMC Bioinformatics* 18:3. doi: 10.1186/s12859-016-1439-1

Meuwissen, T. H., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829. doi: 10.1093/genetics/157.4.1819

Moeinizade, S., Kusmec, A., Hu, G., Wang, L., and Schnable, P. S. (2020). Multi-trait genomic selection methods for crop improvement. *Genetics* 215, 931–945. doi: 10.1534/genetics.120.303305

Morota, G., Abdollahi-Arpanahi, R., Kranis, A., and Gianola, D. (2014). Genome-enabled prediction of quantitative traits in chickens using genomic annotation. *BMC Genomics* 15:109. doi: 10.1186/1471-2164-15-109

Rainer, J., Gatto, L., and Weichenberger, C. X. (2019). ensembldb: an R package to create and use ensembl-based annotation resources. *Bioinformatics* 35, 3151–3153. doi: 10.1093/bioinformatics/btz031

Speed, D., and Balding, D. J. (2014). MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res.* 24, 1550–1557. doi: 10.1101/gr.169375.113

Strandén, I., and Garrick, D. J. (2009). Technical note: derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *J. Dairy Sci.* 92, 2971–2975. doi: 10.3168/jds.2008-1929

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *J. R. Stat. Soc. Ser. B* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Tutar, Y. (2012). Pseudogenes. *Comp. Funct. Genomics* 2012:424526. doi: 10.1155/2012/424526

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi: 10.3168/jds.2007-0980

Wang, X., Xu, Y., Hu, Z., and Xu, C. (2018). Genomic selection methods for crop improvement: current status and prospects. *Crop J.* 6, 330–340. doi: 10.1016/j.cj.2018.03.001

Xu, L., Gao, N., Wang, Z., Xu, L., Liu, Y., Chen, Y., et al. (2020). Incorporating genome annotation into genomic prediction for carcass traits in Chinese simmental beef cattle. *Front. Genet.* 11:481. doi: 10.3389/fgene.2020.00481

Zhang, Z., Ober, U., Erbe, M., Zhang, H., Gao, N., He, J., et al. (2014). Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. *PLoS ONE* 9:e93017. doi: 10.1371/journal.pone.0093017

Zhao, K., Tung, C.-W., Eizenga, G. C., Wright, M. H., Ali, M. L., Price, A. H., et al. (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa. Nat. Commun.* 2:467. doi: 10.1038/ncomms1467

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.