Optimizing genomic prediction model given causal genes in a dairy cattle population

Jinyan Teng,¹ Shuwen Huang,¹ Zitao Chen,¹ Ning Gao,² Shaopan Ye,¹ Shuqi Diao,¹ Xiangdong Ding,³ Xiaolong Yuan,¹ Hao Zhang,¹ Jiaqi Li,¹ and Zhe Zhang^{1*}

¹Guangdong Provincial Key Lab of Agro-Animal Genomics and Molecular Breeding, College of Animal Science, South China Agricultural University, Guangzhou 510642, China

²State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, North Third Road, Guangzhou Higher Education Mega Center, Guangzhou 510006, China

³National Engineering Laboratory for Animal Breeding, Laboratory of Animal Genetics, Breeding and Reproduction, Ministry of Agriculture, College of Animal Science and Technology, China Agricultural University, Beijing 100193, China

ABSTRACT

As genotypic data are moving from SNP chip toward whole-genome sequence, the accuracy of genomic prediction (GP) exhibits a marginal gain, although all genetic variation, including causal genes, are contained in whole-genome sequence data. Meanwhile, genetic analyses on complex traits, such as genome-wide association studies, have identified an increasing number of genomic regions, including potential causal genes, which would be reliable prior knowledge for GP. Many studies have tried to improve the performance of GP by modifying the prediction model to incorporate prior knowledge. Although several plausible results have been obtained from model modification or strategy optimization, most of them were validated in a specific empirical population with a limited variety of genetic architecture for complex traits. An alternative approach is to use simulated genetic architecture with known causal genes (e.g., simulated causative SNP) to evaluate different GP models with given causal genes. Our objectives were to (1) evaluate the performance of GP under a variety of genetic architectures with a subset of known causal genes and (2) compare different GP models modified by highlighting causal genes and different strategies to weight causal genes. In this study, we simulated pseudo-phenotypes under a variety of genetic architectures based on the real genotypes and phenotypes of a dairy cattle population. Besides classical genomic best linear unbiased prediction, we evaluated 3 modified GP models that highlight causal genes as follows: (1) by treating them as fixed effects, (2) by treating them as a separate random component,

and (3) by combining them into the genomic relationship matrix as random effects. Our results showed that highlighting the known causal genes, which explained a considerable proportion of genetic variance in the GP models, increased the predictive accuracy. Combining all given causal genes into the genomic relationship matrix was the optimal strategy under all the scenarios validated, and treating causal genes as a separate random component is also recommended, when more than 20% of genetic variance was explained by known causal genes. Moreover, assigning differential weights to each causal gene further improved the predictive accuracy.

Key words: genomic selection, whole-genome sequence data, causal gene, prior knowledge

INTRODUCTION

Genomic prediction (GP) was first proposed by Meuwissen et al. (2001) and has become a powerful tool employed across the breeding and genetics industry. In practice, GP with low-density SNP chip data has been successfully applied to global dairy cattle breeding (VanRaden et al., 2009). It is ongoing for several other species of animals (Meuwissen et al., 2014; Teng et al., 2019) and plants (Xu et al., 2014; Westhues et al., 2019). Recently, with the rapidly decreasing cost of DNA sequencing (Wetterstrand, 2014) and fast developing genotype imputation techniques, high-density and even whole-genome sequence (WGS) genotype data have become available for major livestock and crop populations. Genetic analysis or breeding practice employing WGS data is thus becoming feasible for more and more species (van Binsbergen et al., 2015; Xu et al., 2018).

Because WGS data includes all genetic variations, theoretically GP using WGS data is expected to achieve a considerable increase in accuracy for complex traits. Although a simulation study reported the advantage of

Received January 20, 2020.

Accepted July 7, 2020.

*Corresponding author: zhezhang@scau.edu.cn

a 40% accuracy improvement when using WGS data (Meuwissen and Goddard, 2010), a marginal increase in accuracy using classical prediction models was observed in a large number of studies on real populations, such as dairy cattle (van Binsbergen et al., 2015), chickens (Ni et al., 2017), and flies (Ober et al., 2012). This marginal increase in accuracy mainly resulted from genotype imputation errors (Brøndum et al., 2014), the limited population size (Druet et al., 2014), and the noise involved in WGS data (van den Berg et al., 2016; Pérez-Enciso et al., 2017). More importantly, using an unoptimized model for GP in such situation is one of the most significant reasons for the marginal increase, because the performance of different GP models varies for different genetic architectures (Daetwyler et al., 2010).

Although the predictive accuracy of GP with WGS data and low-density SNP chip data is basically the same, the power of sequence-based GWAS, which have been used for genetic analyses on complex traits, have demonstrated a considerable increase (Caballero et al., 2015; Frischknecht et al., 2016). As genotypic data are moving from low density to high density or WGS, the number of reported QTL and associations for cattle increased sharply from 13,000 in 2014 to 142,000 in 2020 in the Animal QTL database (Hu et al., 2019). This trend indicated that more reliable QTL (including causal genes) will become available for more economically important traits in the future, though these will be accompanied by an emerging risk of spurious associations. Except for QTL, other biological annotations are also quickly accumulated and can be freely accessed through publicly accessible databases, such as Ensembl (Hunt et al., 2018), the National Center for Biotechnology Information (Agarwala et al., 2018), and the University of California Santa Cruz (UCSC) database (Rosenbloom et al., 2013). Therefore, an increasing amount of prior information (including QTL and biological annotations) was incorporated into GP models for further improvement of the GP accuracy, after the primary exploration of introducing GWAS results in GP (Zhang et al., 2014). This was usually conducted in 2 main ways. On the one hand, different genomic features, such as Gene Ontology terms (Fang et al., 2017), gene annotation (Gao et al., 2018; Rezende et al., 2019), and GWAS results (Zhang et al., 2014), were collected from public data resources as prior information to filter SNP. It should be noted that optimized prior information was required to avoid including too much noise while filtering SNP (Pérez-Enciso et al., 2017). On the other hand, classical GP models were modified to incorporate obtained genomic features, such as treating genomic features as a random com-

ponent (Brøndum et al., 2015) or fixed effects (Lopes et al., 2017) in the GP models. Although several plausible results were obtained with model modification or strategy optimization, most results were validated in a specific empirical population with a limited variety of genetic architecture for complex traits. Given causal genes or quantitative trait nucleotides (QTN), the optimal model of GP is yet to be investigated.

With the increasing availability of population-level WGS data and high-quality annotation information on economically important traits, dairy cattle are potentially the first livestock species for which GP can be conducted considering information of known causal genes. Therefore, focusing on a dairy cattle population, we conducted this study with the aims of (1) evaluating the performance of GP under a variety of genetic architectures with a subset of known causal genes and (2) comparing different GP models modified for highlighting causal genes and different strategies to weight causal genes.

MATERIALS AND METHODS

Population and Data

The German Holstein population used in our study consisted of 2,000 bulls provided by Vereinigte Informationssysteme Tierhaltung w.V. (Zhang et al., 2014). All 2,000 individuals were genotyped with the Illumina Bovine SNP50 BeadChip (54,001 SNP; Matukumalli et al., 2009). Moreover, 770k high-density SNP data of 1,271 animals were obtained from the 1,000 Bull Genome Project (Daetwyler et al., 2014). Taking these 770k high-density SNP data as reference, 54k SNP data of the German Holstein bulls were imputed to 770k via Beagle software (version 5.0; Browning and Browning, 2016). Through genotype imputation, 682,814 SNP were obtained. Quality control of the imputed high-density SNP data was performed with the criteria of minor allele frequency (MAF) $\geq 0.5\%$ and call rate $\geq 95\%$. Finally, 355,699 SNP were retained for further analysis. In addition, conventional EBV with a high reliability for the milk yield (MY) of all 2,000 bulls were available as original phenotypes (y_{org}) in this study.

Simulation of Pseudo-Phenotypes

In dairy cattle, MY is a well-studied complex trait, which has the known major gene *DGAT1* (Grisart et al., 2002), and represents only one complex genetic architecture. To mimic a variety of underlying genetic architectures on such complex traits, we simulated a series of pseudo-phenotypes to evaluate the features of

Figure 1. Schematic diagram of the genetic architecture of real and simulated dairy cattle phenotype. P, A, and R represent phenotypes, additive genetic effects, and residual error, respectively. The component of real phenotype (real) includes the major gene (*DGAT1*, blue), polygenes (red), and residual error (yellow). For the simulated pseudo-phenotype (simulated), the effect of major gene was replaced with simulated causal quantitative trait nucleotides (QTN) using the parameters (parameter) of the bottom row, where N_{QTN} is the numbers of QTN, $\sigma_{QTN_i}^2$ is the total genomic variance of simulated QTN, σ_{DGAT1}^2 is the genomic variance of *DGAT1*, and t is the scale of the genomic variance of simulated causal QTN relative to the genomic variance of *DGAT1*.

different GP models and strategies. The pseudo-phenotypes were simulated by assigning an effect similar to that of *DGAT1* to several SNP, while keeping the existing genetic effects other than those associated with *DGAT1* in the pseudo-phenotypes. In this way, the genetic architecture of pseudo-phenotypes was as real as possible and complex in nature, and its schematic diagram is shown in Figure 1. Based on original phenotypes ($\mathbf{y}_{org}$), the simulation of pseudo-phenotypes ($\mathbf{y}$) was conducted as follows.

First, the substitution effect of all SNP were estimated from the whole population using a linear mixed model via GEMMA software (Zhou and Stephens, 2012), and the estimated marker effects were provided in Supplemental Figure S1 (<https://doi.org/10.3168/jds.2020-18233>). The effect of major gene *DGAT1* (rs109421300, MAF = 0.330) on MY was estimated as -352.437.

Second, a different number of QTN (N_{QTN}) was sampled randomly from 2,000 SNP with the smallest size of estimated effect. These sampled QTN were assigned different effect values to act as true causal genes (i.e., causal QTN) controlling the pseudo-phenotypes.

Then, the causal QTN and effect values were simulated as follow. The total genomic variance of simulated QTN (σ_{QTN}^2) is a function of the genomic variance of *DGAT1* (σ_{DGAT1}^2) and can be expressed as:

$$\sum_{i=1}^{N_{QTN}} \sigma_{QTN_i}^2 = t \times \sigma_{DGAT1}^2, \quad [1]$$

where N_{QTN} is the number of QTN, $\sigma_{QTN_i}^2 = 2p_i(1-p_i)\beta_i^2$, $\sigma_{DGAT1}^2 = 2p(1-p)\alpha_{DGAT1}^2$, in which p_i and p are the MAF, α_{DGAT1} is the substitution effect (-352.437) of *DGAT1*, β_i is the effect value at QTN i sampled from the normal distribution $N(0,1)$. All QTN effect values in vector β are then rescaled to force the summation of all simulated QTN genomic variance to be a function of *DGAT1* genomic variance as in Equation 1. The t is the scale of the simulated σ_{QTN}^2 relative to σ_{DGAT1}^2 , indicating that t times σ_{DGAT1}^2 was assigned to σ_{QTN}^2 for simulating different levels of the effect size of QTN.

Finally, the pseudo-phenotypes were generated by the following formula:

$$\mathbf{y} = \mathbf{y}_{org} - \mathbf{w}\alpha_{DGAT1} + \sum_{i=1}^{N_{QTN}} \mathbf{s}_i \beta_i, \quad [2]$$

where $\mathbf{y}_{org}$ represents the original phenotypes of MY (bulls' EBV), $\mathbf{w}$ is the gene content (coded as 0, 1, or 2) of *DGAT1*, α_{DGAT1} is the substitution effect of *DGAT1*, N_{QTN} is the number of QTN, $\mathbf{s}_i$ is an $N \times 1$ vector of the gene content (coded as 0, 1, or 2) at QTN i .

Parameter Setting. In the preceding procedure (Equation 1), N_{QTN} was assigned ranging from 1, 10, 20, 50, and 100, and t was assigned ranging from 0, 0.1, 0.2, 0.5, 1, 1.5, 2, 5, and 10, where $t = 0$ means without causal QTN related to the simulated trait (i.e., $\sigma_{QTN}^2 = 0$). It should be noted that, when $N_{QTN} = 1$, the rs109421300 (*DGAT1*) was assigned as the causal QTN directly rather than randomly sampled. Under the same N_{QTN} , the effect values (β) of QTN were sampled once, but rescaled according to scale parameter t . Using this approach, traits with different underlying genetic architectures were simulated for further analysis. It should be noted that the true causal genes were assumed to be known for these traits in our study.

Genomic Prediction Models. Our study evaluated 4 GP models including classical genomic BLUP (**GBLUP**) (VanRaden, 2008), GBLUP with fixed effects termed GBLUP_fixed (Lopes et al., 2017), MultiBLUP (Speed and Balding, 2014), and BLUP|GA (Zhang et al., 2014).

Classical GBLUP. Genomic BLUP (VanRaden, 2008) uses information from genomic markers. The model can be expressed as:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}\mathbf{g} + \mathbf{e}, \quad [3]$$

where $\mathbf{y}$ is an $N \times 1$ vector of the observations, $\mathbf{1}$ is the $N \times 1$ vector of ones, μ is the overall mean, $\mathbf{g}$ is a vector of additive genetic values, and $\mathbf{Z}$ is a design matrix that allocates observations to genetic values. It was assumed that $\mathbf{g} \sim N(0, \mathbf{G}\sigma_g^2)$, where σ_g^2 is the additive genetic variance and $\mathbf{G}$ is the marker-based relationship matrix, and $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$, where σ_e^2 is the residual variance and $\mathbf{I}$ is an identity matrix. The $\mathbf{G}$ matrix in this study was constructed with (VanRaden, 2008):

$$\mathbf{G} = \frac{\mathbf{M}\mathbf{D}\mathbf{M}^T}{c}, \quad [4]$$

where $\mathbf{M}$ is a matrix of genotypes and $\mathbf{D}$ is the identity matrix $\mathbf{I}$ in this model; it was defined that $c = 2\sum_{i=1}^m p_i(1-p_i)$, in which p_i is the MAF at SNP i .

GBLUP with Fixed Effects. The GBLUP model with causal QTN modeled as fixed effect (i.e., GBLUP_fixed) is

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}\mathbf{g} + \mathbf{X}\mathbf{b} + \mathbf{e}, \quad [5]$$

where $\mathbf{y}$, $\mathbf{1}$, μ , $\mathbf{Z}$, $\mathbf{g}$, and $\mathbf{e}$ are the same as those in classical GBLUP (Equation 3), $\mathbf{b}$ is a vector of fixed effect values of the causal QTN for the corresponding trait, and $\mathbf{X}$ is the genotype matrix (coded as 0, 1, or 2) of causal QTN.

MultiBLUP. The MultiBLUP with 2 random variance components model can be expressed as

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}\mathbf{g} + \mathbf{Z}\mathbf{g}_u + \mathbf{e}. \quad [6]$$

The identifiers in this model are the same as those in Equation 3, except that $\mathbf{g}_u \sim N(0, \mathbf{S}\sigma_{g_u}^2)$, in which $\mathbf{S}$ is the matrix incorporating causal QTN for the corresponding trait. The $\mathbf{S}$ matrix was constructed via $\mathbf{S} = \frac{\mathbf{M}_1\mathbf{D}\mathbf{M}_1^T}{c_1}$, in which $\mathbf{M}_1$ is the matrix of genotypes for m_1 causal QTN and $\mathbf{D}$ is the identity matrix $\mathbf{I}$ in this model. It was defined that $c_1 = 2\sum_{i=1}^{m_1} p_i(1-p_i)$, in which p_i is the MAF of the causal QTN i .

BLUP|GA. The model expression of the BLUP approach given the genetic architecture (**BLUP|GA**) is the same as that in Equation 3, except that the $\mathbf{G}$ matrix was replaced with the $\mathbf{T}$ matrix integrating the genetic architecture. The $\mathbf{T}$ matrix in this study was constructed according to the literature (Zhang et al., 2014):

$$\mathbf{T} = \frac{\mathbf{M}\mathbf{M}^T + \tau\mathbf{M}_1\mathbf{D}\mathbf{M}_1^T}{c + \tau c_1}, \quad [7]$$

where $\mathbf{M}$ and c are the same as for the $\mathbf{G}$ matrix (Equation 4); $\mathbf{M}_1$ is the genotype matrix of m_1 causal QTN for the corresponding trait; $\mathbf{D}$ is the identity matrix $\mathbf{I}$ in this model; τ is an overall weight factor for m_1 causal QTN; and $c_1 = 2\sum_{i=1}^{m_1} p_i(1-p_i)$, in which p_i is the MAF of the causal QTN i . Equation 7 is equivalent to $\mathbf{T} = \lambda\mathbf{S} + (1-\lambda)\mathbf{G}$ in the previous study (Zhang et al., 2014), in which $\lambda = \frac{\tau c_1}{c + \tau c_1}$ and $\mathbf{S} = \frac{\mathbf{M}_1\mathbf{D}\mathbf{M}_1^T}{c_1}$.

Weighting Strategies for Causal QTN

When causal QTN and their effects were known (i.e., $t > 0$), different weights could be assigned to each QTN to improve the predictive accuracy. To obtain the optimal weighting strategy, 3 scenarios were compared and used to construct the $\mathbf{S}$ matrix for MultiBLUP and BLUP|GA, in which $\mathbf{G}$ matrices were not altered in either model. The major difference is the calculation of diagonal elements of the $\mathbf{D}$ matrix, in which $\mathbf{D} = \text{diag}(h_1, h_2, \dots, h_{m1})$ and $h_1, h_2, \dots, h_{m1}$ are the marker weights. The 3 weighting strategies (i.e., **EQU**, **ABS**, and **SQU**) used in this study were as follows: (1) EQU: assigning equal weights to all QTN, which means that $\mathbf{D}$ is equal to the identity matrix (i.e., $h_i = 1$); (2) ABS: weighting QTN by using the corresponding absolute effect values (i.e., $h_i = |b_i|$) as weight factors; and (3) SQU: weighting QTN by using the squared effect values (i.e., $h_i = \beta_i^2$) as weight factors. It should also be noted that the diagonal of $\mathbf{D}$ was rescaled to make the mean of diagonal elements equal to 1.

Model Assessment

Twenty repetitions of 5-fold cross-validation were conducted to assess the predictive accuracy of all models and strategies. For each repetition, the validation group ($N = 400$) was randomly sampled from the whole population, and the reference group ($N = 1,600$) consisted of the remaining individuals. Variance components were estimated using generalized REML via LDAK software (Speed and Balding, 2019). The predicted genetic values ($\mathbf{Z}\hat{\mathbf{g}}$ for GBLUP, $\mathbf{Z}\hat{\mathbf{g}} + \mathbf{X}\hat{\mathbf{b}}$ for

GBLUP_fixed, $\mathbf{Z}\hat{\mathbf{g}} + \mathbf{Z}\hat{\mathbf{g}}_u$ for MultiBLUP, and $\mathbf{Z}\hat{\mathbf{g}}$ for BLUP|GA) were calculated by solving the mixed model equations for each GP model on the R statistical platform (R Core Team, 2018). The predictive accuracy was defined as the Pearson correlation coefficient between the predicted genetic values and the simulated phenotypic values in the validation group.

RESULTS

Genetic Architectures for Simulated Traits

Using genotypic and phenotypic data from a real dairy cattle population, our study simulated traits with different underlying genetic architectures to explore the optimal GP model and strategy of incorporating causal genes. By varying simulation parameters, we obtained a total of 45 traits (detailed in Supplemental Table S1, <https://doi.org/10.3168/jds.2020-18233>) with different causal QTN numbers (N_{QTN}) or effect sizes and used these in the evaluation of GP models and strategies. Figure 2 shows the proportion of genetic variance explained by either polygenic markers or causal QTN for each simulated trait estimated from the whole data set. Among all the simulated traits, 25 traits had a proportion of genetic variance explained by causal QTN greater than 20%. The estimated proportion of genetic variance explained by causal QTN increased with the increase of simulated genomic variance of causal QTN (σ_{QTN}^2), whereas it decreased with the increase of the QTN number (Figure 2).

Figure 2. Estimated proportion of genetic variance explained by either polygenic markers or causal quantitative trait nucleotides (QTN) for all simulated traits. Title of each frame is the QTN number (N_{QTN}); x-axis indicates the scale (t) of the simulated genomic variance of QTN relative to the genomic variance of $DGAT1$; red/green columns indicate the proportion of total genetic variance explained by polygenes/QTN for each simulated trait. The variance components are estimated by fitting a linear mixed model with 2 random components (polygenes and QTN) using the whole data set.

Figure 3. Accuracy of genomic prediction using different models. The x-axis indicates the scale (t) of the simulated genomic variance of quantitative trait nucleotides (QTN) relative to the genomic variance of $DGAT1$; τ is an overall weight factor for causal QTN in $\mathbf{T}$ matrix of BLUP|GA; and N_{QTN} is the simulated causal QTN number for each trait. The bottom and top of the box represent first and third quartiles; the horizontal line denotes the median; the whiskers correspond to $1.5 \times$ interquartile distance; and dark dots are outliers.

Figure 4. Accuracy of MultiBLUP using different weighting strategies for causal quantitative trait nucleotides (QTN). The x-axis indicates the scale (t) of the simulated genomic variance of QTN relative to the genomic variance of $DGAT1$; N_{QTN} = the simulated causal QTN number for each trait; EQU = assigning equal weights to all QTN; ABS = weighting QTN by using the corresponding absolute effect values as weight factors; SQU = weighting QTN by using the squared effect values as weight factors. The bottom and top of the box represent first and third quartiles; the horizontal line denotes the median; the whiskers correspond to $1.5 \times$ interquartile distance; and dark dots are outliers.

Accuracy of Genomic Prediction Incorporating True Causal QTN

Using traits with different underlying genetic architectures, our study compared the accuracies of 4 GP models (GBLUP, GBLUP_fixed, MultiBLUP, and BLUP|GA). The boxplot of predictive accuracies of all GP models is shown in Figure 3.

Overall, almost all models showed an increase in the predictive accuracy with the increase of σ_{QTN}^2 under the same QTN number. Compared with the 3 prediction models highlighting causal QTN (GBLUP_fixed, MultiBLUP, and BLUP|GA), GBLUP showed less changes (<0.05) in the predictive accuracy with the increase of σ_{QTN}^2 . When QTN had relatively high genomic variance in the tested traits (at $t \geq 1$ for $N_{QTN} =$

1, 10, $t \geq 1.5$ for $N_{QTN} = 20, 50$, and $t \geq 5$ for $N_{QTN} = 100$) or if the proportion of genetic variance explained by QTN was greater than 20% (red line in Figure 2), GBLUP gave the lowest predictive accuracy compared with the other 3 prediction models. Through the comparison of GBLUP and the 2 components models (GBLUP_fixed and MultiBLUP), we found that the 2 components models were worse than GBLUP in terms of their predictive ability for traits with a relatively lower σ_{QTN}^2 (at $t \geq 0.5$ for $N_{QTN} = 1, 10$ and $t \geq 0.2$ for $N_{QTN} = 20$) or if the proportion of genetic variance explained by QTN was less than 20%. Among the 2 components prediction models, MultiBLUP outperformed GBLUP_fixed for 31 out of 45 traits. In all cases, BLUP|GA gave the best performance if using an optimal weight parameter τ , and its advantages over GB-

LUP increased with the increase of σ_{QTN}^2 . The predictive accuracies of BLUP|GA were very similar to MultiBLUP in terms of traits with high σ_{QTN}^2 . In addition, the optimal weight parameter τ increased gradually with the increase of σ_{QTN}^2 , when the BLUP|GA gave the highest predictive accuracy. BLUP|GA yielded a lower predictive accuracy when too large or small values were assigned to the weight parameter τ .

Accuracy of Genomic Prediction Incorporating False Prior Information

In the scenario $t = 0$ (first column in Figure 3), the simulated QTN had zero effect on simulated traits (i.e., no causal QTN), which represents a control group for investigating the accuracy of GP incorporating false prior information (e.g., noncausative SNP). At $t = 0$, GBLUP and BLUP|GA with $\tau = 1$ gave the best predictive ability, whereas the predictive accuracy of BLUP|GA decreased with the increase of the weight parameter τ . Compared with GBLUP and BLUP|GA, GBLUP_fixed and MultiBLUP had a lower predictive accuracy as N_{QTN} increase. Moreover, GBLUP_fixed was worse than MultiBLUP in terms of the predictive ability when incorporating more noncausative SNP into the prediction model.

Accuracy of Genomic Prediction Using Different Weighting Strategies

When causal QTN and their effects were given ($t > 0$), we compared 3 weighting strategies (EQU, ABS, and SQU) for causal QTN in the **S** matrix of MultiBLUP and BLUP|GA, which were detailed above. Figure 4 and Supplemental Figure S2 (<https://doi.org/10.3168/jds.2020-18233>) showed the predictive accuracies of MultiBLUP and BLUP|GA using these 3 weighting strategies to handle causal QTN, respectively. When σ_{QTN}^2 was relatively low, MultiBLUP yielded somewhat different predictive accuracies in different weighting strategies for causal QTN. Additionally, the advantages gained from EQU to ABS to SQU increased with the increase of N_{QTN} . When traits had relatively high σ_{QTN}^2 , we observed small differences in the predictive accuracy between the different weighted strategies. Compared with MultiBLUP, BLUP|GA exhibited some different trends (Supplemental Figure S2) when traits had a high σ_{QTN}^2 , where EQU and ABS yielded the optimal predictive accuracy at $N_{QTN} = 10$ and $N_{QTN} > 10$, respectively.

DISCUSSION

With more and more biological information associated with specific complex traits becoming available in public bioinformatics databases, a comprehensive strategy for incorporating prior knowledge is urgently needed to improve the performance of GP. In the present study, 4 GP models with different ways of utilizing known causal genes were evaluated with simulated phenotypes with a variety of underlying genetic architectures.

Several factors affect the performance of GP, of which the prediction model is an important one in the case where causal genes are given. This has been evidenced by the literature, in which a marginal increase in accuracy was observed for existing GP models, especially for GBLUP (Daetwyler et al., 2010; van Binsbergen et al., 2015; Veerkamp et al., 2016; Al Kalaldehy et al., 2019), in which a flat prior was applied to the distribution of genetic effects across the entire genome. Although Bayesian methods, such as BayesB (Meuwissen et al., 2001) and BayesRc (MacLeod et al., 2016), were proposed to fill the theoretical gap of GBLUP, the computational demand and methodological complexity may hinder their implementation to some extent, especially in breeding practices. Theoretically, the assumption on the genetic architecture of complex traits for GBLUP posed a disadvantage in the case of given causal genes. Although interesting, this is out of the scope of the present study.

To effectively use known causal genes, several model modifications were proposed to overcome the theoretical limitation of GBLUP. Given reasonable prior, 3 modified models were validated in separate conditions (Speed and Balding, 2014; Zhang et al., 2014; Lopes et al., 2017). Meanwhile, they were validated in a unified condition in the present study. Observations from both ours and previous studies support the expectation of models given prior (such as causal genes) provide a better performance. Though not true of all models, at least one modified model highlighting given prior knowledge outperformed GBLUP in all scenarios providing true causal genes (Figure 3). The advantage of modified models over GBLUP increased with the genetic variance explained by known genes (Figure 2). This trend confirmed the theoretical disadvantage of GBLUP in such a condition. Therefore, it is hard to expect a better performance for GBLUP, even when using genome sequence data including all QTN. Inversely, the 3 modified models showed the advantage of utilizing prior knowledge, although they highlighted causal genes in different ways. All these observations and the results of previous studies (Brøndum et al., 2015; Fragomeni et al., 2017; Lopes et al., 2017) together suggested that GP model selection should be a key factor for obtaining

an optimal performance when useful prior is provided. Models highlighting prior information should be given more attention when useful prior is available.

To validate the performance of different ways of highlighting known genes, we tested 3 models, which treated given causal genes as fixed effects (GBLUP_fixed) or a separate random component (MultiBLUP), or combined them into the **G** matrix (BLUP|GA). According to our comparison, the best performance in most tested scenarios was observed in BLUP|GA (Figure 3). This can mainly be attributed to the flexible weight factor (τ) used in the relationship matrix combination. The optimal weight factor increased with the proportion of genetic variance explained by causal QTN. It was also affected by the number of given causal QTN (Figure 3 in our study and Figure 2 in Fragomeni et al., 2017). Though the optimal weight factor in BLUP|GA cannot be determined in advance such as via variance estimation in MultiBLUP, the advantage of BLUP|GA over MultiBLUP suggests that parameter optimization based on the weight factor is worth conducting.

Different from BLUP|GA, MultiBLUP and GBLUP_fixed treat causal QTN as a separate random component or fixed effects. MultiBLUP outperformed GBLUP when the given causal genes explained more than 20% of genetic variance in our simulation study (Figure 2). The results from this study and Pérez-Enciso et al. (2017) together indicate that the proportion of genetic variance explained by causal QTN mainly affects the predictive accuracy. Compared with MultiBLUP, GBLUP_fixed yielded a lower accuracy in most scenarios, especially for scenarios with large N_{QTN} , which indicated that fitting more variables in GBLUP_fixed could not be advantageous though all fitted variable are true causal QTN. But it is obvious that MultiBLUP model is not affected by this factor. Hence, it can be concluded that treating given causal genes as random effects is a better alternative to treat them as fixed effects in GP. When estimating effects of markers in linkage disequilibrium with causal variants, similar observation was reported by Goddard et al. (2009) in a GWAS model.

When the given causal genes are zero in effect (scenario of $t = 0$), false prior information is passed to the GP model. It can be expected that all models highlighting prior knowledge will be affected by false prior knowledge. Our results supported this hypothesis (Figure 3). Therefore, a precondition for the advantage of modified models over GBLUP is that the given prior information (e.g., causal genes) must be true. Accordingly, we concluded that combining given causal genes in the **G** matrix is the optimal strategy when reliable prior knowledge is provided. Treating given causal genes as a separate random component is also recom-

mended when the genetic variance explained by causal genes exceeds a certain proportion.

To further optimize the strategy of utilizing prior knowledge, situations including both given causal genes and their effect sizes (the relative importance) were evaluated for BLUP|GA and MultiBLUP. We evaluated 3 strategies of using the effect size of given genes, denoted as EQU, ABS, and SQU. The strategy of treating causal QTN with ABS gave the best performance when causal QTN explained a high proportion of genetic variance, which is a rare case in practice. In scenarios of multi-QTN explaining a lower proportion of genetic variance, differential weighting strategies (i.e., SQU and ABS) showed a higher accuracy compared with the unweighting strategy (i.e., EQU). All these results and those of previous studies (Zhang et al., 2016; Teissier et al., 2019) together indicate that assigning special weights to each QTN can yield a better performance in GP when precise QTN effect sizes are known.

In this study, we evaluated several modified models and strategies for incorporating prior knowledge, under the assumption that the prior knowledge is provided without error. One question that arises in practice relates to how to effectively obtain reliable prior information from sequence data or public data sets. One way is obtaining high-quality annotation information from publicly released databases such as Animal QTLdb, Ensembl, and the National Center for Biotechnology Information. Such databases provided abundant biological knowledge related to traits of interest including the QTL region, gene function, or biological pathway. Their availabilities and effectiveness have been reported by several studies (Zhang et al., 2014; Fang et al., 2017; Gao et al., 2017). Another way is to detect potential causal variants from a larger population such as across breeds or admixed populations (Hayes et al., 2014; van den Berg et al., 2016). No matter which information source used, it should not be ignored that problems of false positive and unvalidated results or imprecise priors exist in current practice. Hence, strategy of enhancing signal and reducing noise should be give more attention.

Another question that might arise in practice relates to how to link our simulated situation to the practical situation. To mimic the practical situation and facilitate model comparison, a pseudo-phenotype was generated by combining the true EBV and simulated genetic effect from known genes in our study. The pseudo-phenotype includes 2 components: a real component and a simulated component. The real component with polygenic underlying genetic architecture was generated by removing the effect of the real major gene *DGAT1* from the original EBV. Additionally, the second component includes simulated genes with

a variety of genetic architectures (number of genes and effect size of genes, and overall heritability). With this design, we mimicked the practical situation where a set of known causal genes for a complex trait was treated as prior knowledge and other unknown genes were also included in the genome data. Based on our current understanding on the genetic architecture of complex traits, the scenario with a low proportion of genetic variance explained by causal QTN (Figure 2) would be more realistic, because a large part of genetic variance may be explained by low-frequency SNP that are hard to find in a selective case livestock population (Pérez-Enciso et al., 2017). However, the proportion of genetic variance explained by known QTN is expected to increase with more details of genetic architecture uncovered by genetic dissection work in the near future. More and more genetic variance will be captured by the increasingly available genetic variants or omics data (Yang et al., 2015; Xiang et al., 2019). It should be noted that the genetic architecture of complex traits is trait-specific. It is hard to determine which scenario in our simulation is the closest one to reality. Therefore, we have provided a spectrum of genetic architecture, including the number and the explained genetic variance of the simulated QTN, for reference (Figure 2). In practice, one can decide which scenario is the closest one to one's own reality, when a set of genes (prior knowledge, such as a pathway) is obtained and their explained genetic variance is estimated.

In this study, random cross-validation was used to evaluate different prediction models and strategies. Overestimation of the prediction accuracy may exist because selection or the relationship between individuals was not accounted for. The main purpose of the present study was to compare different models, rather than report the accuracy of GP in this cattle population. Conducting a fair comparison with reliable results was thus more important. Therefore, for all tested scenarios with comparable parameters, the training and validation of individuals were the same for all tested models within each replication, which made the comparison fair. Practically, the prediction accuracy in a livestock population might be evaluated via forward prediction to avoid overestimation resulting from random cross-validation.

CONCLUSIONS

Under the assumption that causal genes were given, we evaluated 4 GP models with a broad spectrum of simulated genetic architectures. Due to the uninformative assumption on the distribution of the genetic effect across the genome, the accuracy of GBLUP is hard to benefit from including causal genes. All mod-

els highlighting the known causal genes could increase the accuracy of GP compared with the model under a flat prior, e.g., GBLUP. Combining all given causal genes into the **G** matrix is the optimal strategy for all scenarios validated, and treating causal genes as a separate random component is also recommended although a certain proportion of genetic variance (over 20% in our simulation) was explained by known causal genes. Moreover, if the genetic contribution of each causal gene is given, assigning differential weights for each causal gene is suggested to further improve the accuracy of GP. In practice, it should be noted that the genetic architecture of the trait of interest and the features of known causal genes should be comprehensively considered to select the optimal GP model.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Beijing; 31772556, 31671327), the Earmarked Fund for China Agriculture Research System (CARS-35), the Local Innovative and Research Teams Project of Guangdong Province (2019BT02N630), and the National Supercomputer Center in Guangzhou. We thank the Vereinigte Informationssysteme Tierhaltung w.V. for providing the German Holstein data. We also acknowledge and appreciate the constructive critiques provided by 2 anonymous reviewers and the editor, which improved this manuscript significantly. The authors declare that they have no competing interests.

REFERENCES

- Al Kalaldehy, M., J. Gibson, N. Duijvesteijn, H. D. Daetwyler, I. MacLeod, N. Moghaddar, S. H. Lee, and J. H. J. van der Werf. 2019. Using imputed whole-genome sequence data to improve the accuracy of genomic prediction for parasite resistance in Australian sheep. *Genet. Sel. Evol.* 51:32. <https://doi.org/10.1186/s12711-019-0476-4>.
- Brøndum, R. F., B. Guldbrandtsen, G. Sahana, M. Lund, and G. Su. 2014. Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. *BMC Genomics* 15:728. <https://doi.org/10.1186/1471-2164-15-728>.
- Brøndum, R. F., G. Su, L. Janss, G. Sahana, B. Guldbrandtsen, D. Boichard, and M. S. Lund. 2015. Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. *J. Dairy Sci.* 98:4107–4116. <https://doi.org/10.3168/jds.2014-9005>.
- Browning, B. L., and S. R. Browning. 2016. Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.* 98:116–126. <https://doi.org/10.1016/j.ajhg.2015.11.020>.
- Caballero, A., A. Tenesa, and P. D. Keightley. 2015. The nature of genetic variation for complex traits revealed by GWAS and regional heritability mapping analyses. *Genetics* 201:1601–1613. <https://doi.org/10.1534/genetics.115.177220>.
- Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. van Binsbergen, R. F. Brøndum, X. Liao, A. Djari, S. C. Rodriguez, C. Grohs, D. Esquerré, O. Bouchez, M.-N. Rossignol, C. Klopp, D. Rocha, S. Fritz, A. Eggen, P. J. Bowman, D. Coote, A. J. Chamberlain, C.

- Anderson, C. P. VanTassell, I. Hulsege, M. E. Goddard, B. Guldbrandtsen, M. S. Lund, R. F. Veerkamp, D. A. Boichard, R. Fries, and B. J. Hayes. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat. Genet.* 46:858–865. <https://doi.org/10.1038/ng.3034>.
- Daetwyler, H. D., R. Pong-Wong, B. Villanueva, and J. A. Woolliams. 2010. The impact of genetic architecture on genome-wide evaluation methods. *Genetics* 185:1021–1031. <https://doi.org/10.1534/genetics.110.116855>.
- Druet, T., I. M. Macleod, and B. J. Hayes. 2014. Toward genomic prediction from whole-genome sequence data: Impact of sequencing design on genotype imputation and accuracy of predictions. *Heredity* 112:39–47. <https://doi.org/10.1038/hdy.2013.13>.
- Fang, L., G. Sahana, P. Ma, G. Su, Y. Yu, S. Zhang, M. S. Lund, and P. Sørensen. 2017. Use of biological priors enhances understanding of genetic architecture and genomic prediction of complex traits within and between dairy cattle breeds. *BMC Genomics* 18:604. <https://doi.org/10.1186/s12864-017-4004-z>.
- Fragomeni, B. O., D. A. L. Lourenco, Y. Masuda, A. Legarra, and I. Misztal. 2017. Incorporation of causative quantitative trait nucleotides in single-step GBLUP. *Genet. Sel. Evol.* 49:59. <https://doi.org/10.1186/s12711-017-0335-0>.
- Frischknecht, M., H. Signer-Hasler, T. Leeb, S. Rieder, and M. Neuditschko. 2016. Genome-wide association studies based on sequence-derived genotypes reveal new QTL associated with conformation and performance traits in the Franches-Montagnes horse breed. *Anim. Genet.* 47:227–229. <https://doi.org/10.1111/age.12406>.
- Gao, N., J. W. R. Martini, Z. Zhang, X. Yuan, H. Zhang, H. Simianer, and J. Li. 2017. Incorporating gene annotation into genomic prediction of complex phenotypes. *Genetics* 207:489–501. <https://doi.org/10.1534/genetics.117.300198>.
- Gao, N., J. Teng, S. Ye, X. Yuan, S. Huang, H. Zhang, X. Zhang, J. Li, and Z. Zhang. 2018. Genomic prediction of complex phenotypes using genetic similarity based relatedness matrix. *Front. Genet.* 9:364. <https://doi.org/10.3389/fgene.2018.00364>.
- Goddard, M. E., N. R. Wray, K. Verbyla, and P. M. Visscher. 2009. Estimating effects and making predictions from genome-wide marker data. *Stat. Sci.* 24:517–529. <https://doi.org/10.1214/09-STS306>.
- Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell. 2002. Positional candidate cloning of a QTL in dairy cattle: Identification of a missense mutation in the bovine *DGAT1* gene with major effect on milk yield and composition. *Genome Res.* 12:222–231. <https://doi.org/10.1101/gr.224202>.
- Hayes, B. J., I. M. MacLeod, H. D. Daetwyler, P. J. Bowman, A. J. Chamberlain, C. J. Vander Jagt, A. Capitan, H. Pausch, P. Stothard, X. Liao, C. Schrooten, E. Mullaart, R. Fries, B. Guldbrandtsen, M. S. Lund, D. A. Boichard, R. F. Veerkamp, C. P. VanTassell, B. Gredler, T. Druet, A. Bagnato, J. Vilkkilä, D. J. DeKoning, E. Santus, and M. E. Goddard. 2014. Genomic prediction from whole genome sequence in livestock: The 1000 Bull Genomes Project. Page in Proceedings of the 10th World Congress on Genetics Applied to Livestock Production, Vancouver, Canada.
- Hu, Z.-L., C. A. Park, and J. M. Reecy. 2019. Building a livestock genetic and genomic information knowledgebase through integrative developments of Animal QTLdb and CorrDB. *Nucleic Acids Res.* 47(D1):D701–D710. <https://doi.org/10.1093/nar/gky1084>.
- Hunt, S. E., W. McLaren, L. Gil, A. Thormann, H. Schuilenburg, D. Sheppard, A. Parton, I. M. Armean, S. J. Trevanion, P. Flicek, and F. Cunningham. 2018. Ensembl variation resources. Database (Oxford) 2018. <https://doi.org/10.1093/database/bay119>.
- Lopes, M. S., H. Bovenhuis, M. van Son, Ø. Nordbø, E. H. Grindflek, E. F. Knol, and J. W. M. Bastiaansen. 2017. Using markers with large effect in genetic and genomic predictions. *J. Anim. Sci.* 95:59. <https://doi.org/10.2527/jas.2016.0754>.
- MacLeod, I. M., P. J. Bowman, C. J. Vander Jagt, M. Haile-Mariam, K. E. Kemper, A. J. Chamberlain, C. Schrooten, B. J. Hayes, and M. E. Goddard. 2016. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC Genomics* 17:144. <https://doi.org/10.1186/s12864-016-2443-6>.
- Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton, J. O'Connell, S. S. Moore, T. P. L. Smith, T. S. Sonstegard, and C. P. Van Tassell. 2009. Development and characterization of a high density SNP genotyping assay for cattle. *PLoS One* 4:e5350. <https://doi.org/10.1371/journal.pone.0005350>.
- Meuwissen, T., and M. Goddard. 2010. Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics* 185:623–631. <https://doi.org/10.1534/genetics.110.116590>.
- Meuwissen, T.H., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829.
- Meuwissen, T. H. E., J. Odegard, I. Andersen-Ranberg, and E. Grindflek. 2014. On the distance of genetic relationships and the accuracy of genomic prediction in pig breeding. *Genet. Sel. Evol.* 46:49. <https://doi.org/10.1186/1297-9686-46-49>.
- NCBI Resource Coordinators. 2018. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 46(D1):D8–D13. <https://doi.org/10.1093/nar/gkx1095>.
- Ni, G., D. Caverio, A. Fangmann, M. Erbe, and H. Simianer. 2017. Whole-genome sequence-based genomic prediction in laying chickens with different genomic relationship matrices to account for genetic architecture. *Genet. Sel. Evol.* 49:8. <https://doi.org/10.1186/s12711-016-0277-y>.
- Ober, U., J. F. Ayroles, E. A. Stone, S. Richards, D. Zhu, R. A. Gibbs, C. Stricker, D. Gianola, M. Schlather, T. F. C. Mackay, and H. Simianer. 2012. Using whole-genome sequence data to predict quantitative trait phenotypes in *Drosophila melanogaster*. *PLoS Genet.* 8:e1002685. <https://doi.org/10.1371/journal.pgen.1002685>.
- Pérez-Enciso, M., N. Forneris, G. de los Campos, and A. Legarra. 2017. Evaluating sequence-based genomic prediction with an efficient new simulator. *Genetics* 205:939–953. <https://doi.org/10.1534/genetics.116.194878>.
- R Core Team. 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- Rezende, F. M., J. P. Nani, and F. Peñagaricano. 2019. Genomic prediction of bull fertility in US Jersey dairy cattle. *J. Dairy Sci.* 102:3230–3240. <https://doi.org/10.3168/jds.2018-15810>.
- Rosenbloom, K. R., C. A. Sloan, V. S. Malladi, T. R. Dreszer, K. Learned, V. M. Kirkup, M. C. Wong, M. Maddren, R. Fang, S. G. Heitner, B. T. Lee, G. P. Barber, R. A. Harte, M. Diekhans, J. C. Long, S. P. Wilder, A. S. Zweig, D. Karolchik, R. M. Kuhn, D. Haussler, and W. J. Kent. 2013. ENCODE data in the UCSC Genome Browser: Year 5 update. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gks1172>.
- Speed, D., and D. J. Balding. 2014. MultiBLUP: Improved SNP-based prediction for complex traits. *Genome Res.* 24:1550–1557. <https://doi.org/10.1101/gr.169375.113>.
- Speed, D., and D. J. Balding. 2019. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nat. Genet.* 51:277–284. <https://doi.org/10.1038/s41588-018-0279-5>.
- Teissier, M., H. Larroque, and C. Robert-Granie. 2019. Accuracy of genomic evaluation with weighted single-step genomic best linear unbiased prediction for milk production traits, udder type traits, and somatic cell scores in French dairy goats. *J. Dairy Sci.* 102:3142–3154. <https://doi.org/10.3168/jds.2018-15650>.
- Teng, J., N. Gao, H. Zhang, X. Li, J. Li, H. Zhang, X. Zhang, and Z. Zhang. 2019. Performance of whole genome prediction for growth traits in a crossbred chicken population. *Poult. Sci.* 98:1968–1975. <https://doi.org/10.3382/ps/pey604>.
- van Binsbergen, R., M. P. L. Calus, M. C. A. M. Bink, F. A. van Eeuwijk, C. Schrooten, and R. F. Veerkamp. 2015. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. *Genet. Sel. Evol.* 47:71. <https://doi.org/10.1186/s12711-015-0149-x>.
- van den Berg, I., D. Boichard, B. Guldbrandtsen, and M. S. Lund. 2016. Using sequence variants in linkage disequilibrium with causative mutations to improve across-breed prediction in dairy cattle:

- A simulation study. *G3* (Bethesda) 6:2553–2561. <https://doi.org/10.1534/g3.116.027730>.
- VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91:4414–4423. <https://doi.org/10.3168/jds.2007-0980>.
- VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.* 92:16–24. <https://doi.org/10.3168/jds.2008-1514>.
- Veerkamp, R. F., A. C. Bouwman, C. Schrooten, and M. P. L. Calus. 2016. Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein-Friesian cattle. *Genet. Sel. Evol.* 48:95. <https://doi.org/10.1186/s12711-016-0274-1>.
- Westhues, M., C. Heuer, G. Thaller, R. Fernando, and A. E. Melchinger. 2019. Efficient genetic value prediction using incomplete omics data. *Theor. Appl. Genet.* 132:1211–1222. <https://doi.org/10.1007/s00122-018-03273-1>.
- Wetterstrand, K. A. 2014. DNA sequencing costs: Data. National Human Genome Research Institute. Accessed Jan. 20, 2020. www.genome.gov/sequencingcostsdata.
- Xiang, R., I. van den Berg, I. M. MacLeod, B. J. Hayes, C. P. Prowse-Wilkins, M. Wang, S. Bolormaa, Z. Liu, S. J. Rochfort, C. M. Reich, B. A. Mason, C. J. Vander Jagt, H. D. Daetwyler, M. S. Lund, A. J. Chamberlain, and M. E. Goddard. 2019. Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. *Proc. Natl. Acad. Sci. USA* 116:19398–19408. <https://doi.org/10.1073/pnas.1904159116>.
- Xu, S., D. Zhu, and Q. Zhang. 2014. Predicting hybrid performance in rice using genomic best linear unbiased prediction. *Proc. Natl. Acad. Sci. USA* 111:12456–12461. <https://doi.org/10.1073/pnas.1413750111>.
- Xu, Y., X. Wang, X. Ding, X. Zheng, Z. Yang, C. Xu, and Z. Hu. 2018. Genomic selection of agronomic traits in hybrid rice using an NCII population. *Rice* (NY) 11:32. <https://doi.org/10.1186/s12284-018-0223-4>.
- Yang, J., A. Bakshi, Z. Zhu, G. Hemani, A. A. E. Vinkhuyzen, S. H. Lee, M. R. Robinson, J. R. B. Perry, I. M. Nolte, J. V. Van Vliet-Ostaptchouk, H. Snieder, T. Esko, L. Milani, R. Mägi, A. Metspalu, A. Hamsten, P. K. E. Magnusson, N. L. Pedersen, E. Ingelsson, N. Soranzo, M. C. Keller, N. R. Wray, M. E. Goddard, and P. M. Visscher. 2015. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* 47:1114–1120. <https://doi.org/10.1038/ng.3390>.
- Zhang, X., D. Lourenco, I. Aguilar, A. Legarra, and I. Misztal. 2016. Weighting strategies for single-step genomic BLUP: An iterative approach for accurate calculation of GEBV and GWAS. *Front. Genet.* 7:151. <https://doi.org/10.3389/fgene.2016.00151>.
- Zhang, Z., U. Ober, M. Erbe, H. Zhang, N. Gao, J. He, J. Li, and H. Simianer. 2014. Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. *PLoS One* 9:e93017. <https://doi.org/10.1371/journal.pone.0093017>.
- Zhou, X., and M. Stephens. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44:821–824. <https://doi.org/10.1038/ng.2310>.

ORCIDS

Jinyan Teng <https://orcid.org/0000-0002-3046-1452>
 Xiangdong Ding <https://orcid.org/0000-0001-8949-6206>
 Zhe Zhang <https://orcid.org/0000-0001-7338-7718>