



Genomic Prediction of Complex Phenotypes Using Genic Similarity Based Relatedness Matrix

Ning Gao, Jinyan Teng, Shaopan Ye, Xiaolong Yuan, Shuwen Huang, Hao Zhang, Xiquan Zhang, Jiaqi Li* and Zhe Zhang*

National Engineering Research Center for Breeding Swine Industry, Guangdong Provincial Key Lab of Agro-Animal Genomics and Molecular Breeding, College of Animal Science, South China Agricultural University, Guangzhou, China

OPEN ACCESS

Edited by:

Luis Varona,
Universidad de Zaragoza, Spain

Reviewed by:

Francisco Peñagaricano,
University of Florida, United States
Evangelina López De Maturana,
Centro Nacional de Investigaciones
Oncológicas CNIO-Fundación Cáncer
FUCA, Spain

*Correspondence:

Jiaqi Li
jqli@scau.edu.cn
Zhe Zhang
zhezhang@scau.edu.cn

Specialty section:

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

Received: 24 May 2018

Accepted: 21 August 2018

Published: 31 August 2018

Citation:

Gao N, Teng J, Ye S, Yuan X,
Huang S, Zhang H, Zhang X, Li J and
Zhang Z (2018) Genomic Prediction of
Complex Phenotypes Using Genic
Similarity Based Relatedness Matrix.
Front. Genet. 9:364.
doi: 10.3389/fgene.2018.00364

In the last years, a series of methods for genomic prediction (GP) have been established, and the advantages of GP over pedigree best linear unbiased prediction (BLUP) have been reported. However, the majority of previously proposed GP models are purely based on mathematical considerations while seldom take the abundant biological knowledge into account. Prediction ability of those models largely depends on the consistency between the statistical assumptions and the underlying genetic architectures of traits of interest. In this study, gene annotation information was incorporated into GP models by constructing haplotypes with SNPs mapped to genic regions. Haplotype allele similarity between pairs of individuals was measured through different approaches at single gene level and then converted into whole genome level, which was then treated as a special kernel and used in kernel based GP models. Results shown that the gene annotation guided methods gave higher or at least comparable predictive ability in some traits, especially in the Arabidopsis dataset and the rice breeding population. Compared to SNP models and haplotype models without gene annotation, the gene annotation based models improved the predictive ability by 0.56~26.67% in the Arabidopsis and 1.62~16.53% in the rice breeding population, respectively. However, incorporating gene annotation slightly improved the predictive ability for several traits but did not show any extra gain for the rest traits in a chicken population. In conclusion, integrating gene annotation into GP models could be beneficial for some traits, species, and populations compared to SNP models and haplotype models without gene annotation. However, more studies are yet to be conducted to implicitly investigate the characteristics of these gene annotation guided models.

Keywords: genomic prediction, genomic selection, gene annotation, haplotype models, complex phenotypes

INTRODUCTION

Genomic prediction (GP) (Meuwissen et al., 2001) is a powerful tool in the fields of plant and animal breeding and human complex traits and disease risk prediction. In the past decade, a series of GP approaches have been proposed, including the marker effect methods (Meuwissen et al., 2001; Habier et al., 2011; Gianola, 2013) and genomic best linear unbiased prediction (GBLUP) (VanRaden, 2008).

Currently, standard GP models estimate marker effects and calculate individual genetic values via statistical models, but most of them pay less attention to the underlying connection between the complex genetic architecture and the often simplistic mathematical formulas. Reviewing the literatures and the biological databases, abundant of biological knowledge about trait genetic architecture, gene function, regulation patterns, and gene interaction networks have been quickly accumulated. The potential usefulness of biological knowledge to accelerate GP models has been illustrated by several studies (Zhang et al., 2014; Edwards et al., 2016; Gao et al., 2017). However, the questions about what kind of biological knowledge can be used, how to integrate the prior knowledge into GP models, and how much extra predictive ability can be obtained from the assisted information still need more investigations. Under GBLUP framework, Zhang et al. (2014) incorporated the previously reported quantitative trait loci (QTLs) collected in the animal QTLdb (<http://www.animalgenome.org/QTLdb>) (Hu et al., 2016) into genomic prediction model, where markers were weighted according to the frequency of corresponding genomic regions being reported likely containing QTL when constructing genomic relatedness matrix. Through this way, different variances were assumed among genomic regions and predictive ability was improved, especially for traits controlled by large effect genes (Zhang et al., 2014). Similarly, $-\log_{10}(p)$, where p was the p -value for a marker on the outcomes of interest, was utilized into genomic prediction by weighting SNPs according to $-\log_{10}(p)$ when constructing relatedness matrices and through which predictive ability was enhanced (de Los Campos et al., 2013; Ramstein et al., 2016). In a Bayesian model, instead of using an uniform π (the proportion of markers with zero effects) for all markers, Gao et al. (2015) transferred the GWAS p -values into a locus-specific π and used this genetic architecture derived π into the genomic prediction model. The predictive ability of BayesB was improved by the locus-specific π . In some of the latest publications, more types of biological knowledge were incorporated into genomic prediction by partitioning markers into classes based on their functional annotation (Morota et al., 2014; Do et al., 2015; Abdollahi-Arpanahi et al., 2016; MacLeod et al., 2016) or gene ontology categories (Edwards et al., 2016; Abdollahi-Arpanahi et al., 2017).

Compared to the pedigree BLUP (Henderson, 1975), SNP based GP models (Meuwissen et al., 2001; VanRaden, 2008) show higher predictive ability under many circumstances. In both breeding value models and marker effect models, the underlying mechanism of GP was tracing QTL effects through dense genetic markers (usually SNPs) that were in linkage disequilibrium with the potential neighbor QTLs. However, on one hand, for genes or QTLs harboring more than two alleles, the bi-allelic SNP might not be adequate for tracing the multi-allelic gene effects. On the other hand, in the breeding value GP model, the SNP derived relatedness to some extent reflect the IBS (Identity by state) rather than IBD (Identity by descent). Even though the haplotypes can neither reflect the IBD perfectly, alleles from the same haplotype are more likely to be IBD. Thus, an alternative way to the existing models is using the multi-allelic genotypes in GP by constructing haploblocks with consecutive SNPs. The benefit GP gained from

haplotype models has been shown in several studies (Calus et al., 2008; Meuwissen et al., 2014; Cuyabano et al., 2015; Da, 2015; Gao et al., 2017).

In several previously proposed haplotype based GP models, “artificial markers” were constructed for each haplotype allele, and relatedness matrix was constructed by matrix product of the artificial marker matrix (Calus et al., 2008; Meuwissen et al., 2014; Cuyabano et al., 2015; Da, 2015; Gao et al., 2017), or categorical models were introduced for modeling the haplotype effects (Gao et al., 2017). Alternatively, haplotype based relatedness matrix could be built by firstly calculating a haplotype allele similarity matrix for each haploblock and then converting the allele similarity matrix into individual similarity matrix (Hickey et al., 2013). From the aspect of kernel regression (Gianola et al., 2006; Gianola and van Kaam, 2008), the similarity matrix could be treated as a specific kernel and used in GP in the framework of kernel regression.

In the haplotype models, haploblocks can be defined by considering the linkage disequilibrium among a set of consecutive SNPs (Calus et al., 2008; Cuyabano et al., 2015; Da, 2015) or the number of haplotype alleles in certain haploblock (Meuwissen et al., 2014). Recently, with the aim of defining predictors according to known functioning units, Gao et al. (2017) proposed a strategy to incorporate gene annotation into GP by restricting the haploblock to the protein coding regions. Though the predictive ability of GP models were improved by defining haplotypes according to the structural genes in many complex traits (Gao et al., 2017), more alternative approaches for building genic relatedness matrices need to be examined in order to provide more choices and gain much extra predictive ability. In this study, we (1) constructed haplotypes in the protein coding gene regions, (2) calculated genomic relatedness matrix by firstly constructing haplotype similarity matrices and then converting them into individual similarity matrices, and (3) performed GP utilizing the genic haplotype relatedness matrix. Technically, a haplotype allele similarity matrix was calculated within each haplotype block and converted into individual similarity matrix. GP was performed under the kernel regression framework by treating the individual similarity matrix as a certain kernel.

MATERIALS AND METHODS

In order to build haploblocks in genic regions, SNPs were mapped to protein coding genes according to their corresponding physical positions. For each gene, haplotypes were constructed throughout the gene under consideration. Within each haplotype block, allele similarity matrix was constructed by considering the SNP matching pattern between haplotype alleles. Furthermore, the allele similarity matrix was converted into individual similarity matrix. The final relatedness matrix was calculated by averaging the similarity matrices for all haploblocks. Finally, the genic haplotype similarity based relatedness matrix was used for GP. Three populations of rice, Arabidopsis, and yellow chicken were utilized for model validation (Table 2). We would explain these procedures in the following sections.

Mapping SNPs to Pathways

The latest version of the gene annotation of each considered species was downloaded from Ensemble (<http://www.ensembl.org>) using the biomaRt package (Durinck et al., 2005, 2009) of the R statistical platform (R Development Core Team, 2016) (Table 2). Only genes indicated as “protein_coding” by the “gene_biotype” attribute were considered. Gene boundaries were extended by 5 kb in both upstream and downstream flanking regions to include possible regulatory elements. SNPs that were available for GP were mapped to these genic regions based on their corresponding physical positions. After the SNP mapping step, SNP sets were formed for genes with at least one mapped marker. For genes with only one mapped SNP, the corresponding haplotype block existed of only this marker. For genes with more than one mapped SNPs, phased alleles of the corresponding SNPs were combined into haplotypes with the approach described by Meuwissen et al. (2014). Briefly, haplotypes were built via the following steps.

Initialization: For each gene, start with the first SNP $j = 1$.

Step 1: Include SNP $j + 1$ into the haploblock.

Step 2: Determine the number of alleles of the haploblock defined by these $j + 1$ markers across the whole population.

Step 3: Repeat step 1 and step 2 if the number of alleles remained below a previously chosen threshold restricting the number of alleles of a haploblock (we used 10 as proposed by Meuwissen et al., 2014). Otherwise, if the number of alleles exceeded this threshold, the lastly added SNP was excluded from the current haploblock and used as the starting position of the next haploblock. Return the alleles of the current haploblock and go to the initialization step with the lastly added SNP to define the next haploblock. Repeat this procedure until all SNPs on the currently considered gene were processed.

This approach produced one or more haploblocks with at least two haplotype alleles per block for each gene. Subsequently, the genic similarity matrix could be constructed using these haplotypes.

Genic Similarity Based Relatedness Matrices

Hickey et al. (2013) introduced three approaches for constructing haplotype allele similarity matrices. In the first strategy, similarity between pairs of haplotype alleles were measured as the proportion of matched loci in current haploblock. The second strategy took not only the proportion of matched loci, but also the length of matched segments into consideration. For more details about those two approaches, please refer to the next sections. Moreover, the allele frequencies were further considered in a third strategy. The former two strategies were used in this study to construct allele similarity matrices and further convert into individual similarity matrices. The third approach was not used in the present study. Because its performance was not better than others, and it needs to use the allele frequency, which could not be estimated accurately from small populations. In the following, we illustrated the procedures for calculating allele similarity and individual similarity matrices with a small example. Table 1 showed the genotypes of five individuals and 10 consecutive SNPs from a certain gene. The SNP genotypes were phased and four different haplotype alleles were defined by these markers.

The first strategy calculated haplotype similarity by counting the number of matched SNPs between haplotypes and dividing by the total number of markers contained in the haplotype. In a formula form, the haplotype similarity score was calculated as $h_1 = \frac{n_s}{N}$, where h_1 was the similarity score, n_s was the number of matched SNPs between two haplotypes, and N was the number of SNPs in current haplotype block. For example, *hap1* and *hap2* in Table 1 shared the same SNP alleles for markers *M2*, *M4*, and *M10*. The similarity between *hap1* and *hap2* was calculated as $3/10 = 0.3$. The similarity score between a haplotype and itself equaled to 1. Therefore, similarity matrix of the four haplotypes shown in Table 1 was calculated as H_1 .

$$H_1 = \begin{bmatrix} 1 & 0.3 & 0.4 & 0.3 \\ 0.3 & 1 & 0.3 & 0.8 \\ 0.4 & 0.3 & 1 & 0.5 \\ 0.3 & 0.8 & 0.5 & 1 \end{bmatrix}$$

TABLE 1 | Genotype matrix of five individuals and 10 consecutive markers from a certain protein coding gene.

| Individuals | Gamete | Haplotypes | SNPs mapped to gene | | | | | | | | | |
|-------------|----------|--------------|---------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|------------|
| | | | <i>M1</i> | <i>M2</i> | <i>M3</i> | <i>M4</i> | <i>M5</i> | <i>M6</i> | <i>M7</i> | <i>M8</i> | <i>M9</i> | <i>M10</i> |
| <i>id1</i> | Paternal | <i>hap 1</i> | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 |
| | Maternal | <i>hap 4</i> | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| <i>id2</i> | Paternal | <i>hap 2</i> | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| | Maternal | <i>hap 1</i> | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 |
| <i>id3</i> | Paternal | <i>hap 3</i> | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 |
| | Maternal | <i>hap 4</i> | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| <i>id4</i> | Paternal | <i>hap 2</i> | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| | Maternal | <i>hap 2</i> | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| <i>id5</i> | Paternal | <i>hap 2</i> | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| | Maternal | <i>hap 3</i> | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 |

A haplotype block contains four haplotype alleles is defined by these 10 consecutive markers from a protein coding gene.

In the second strategy, the measurement of haplotype similarity took the length of matching segments into account, where the similarity score increased as the number of consecutive matching SNPs increased. For certain pairs of haplotypes, the final similarity was the sum of all matched segments. Within each segment, the similarity score was calculated as the squared numbers of matching SNPs. Segments containing only one matching SNP was scored one. The overall similarity scores were further standardized by dividing the scores by the maximum of the similarity scores and taking the square root to ensure values with the scale of [0,1]. In a formula form, the haplotype similarity score was calculated as $h_2 = \sqrt{\frac{\sum_{l=1}^L n_{sl}^2}{N^2}}$, where L was the number of matched segments between pairs of haplotypes, n_{sl} was the number of matched SNPs in the l th segment, and N was the number of SNPs in current haplotype block. For example, *hap2* and *hap4* in **Table 1** shared two matching segments, the first segment contained one marker (M1) and the second segment contained seven markers (M4~M10). The similarity scores of the two segments were $1^2 = 1$ and $7^2 = 49$, respectively. Therefore, the final similarity between *hap2* and *hap4* was $\sqrt{(1 + 49)/100} = 0.71$. The similarity matrix of the four haplotypes in **Table 1** was represented in \mathbf{H}_2 .

$$\mathbf{H}_2 = \begin{bmatrix} 1 & 0.17 & 0.24 & 0.22 \\ 0.17 & 1 & 0.30 & 0.71 \\ 0.24 & 0.30 & 1 & 0.36 \\ 0.22 & 0.71 & 0.36 & 1 \end{bmatrix}$$

For comparison, a third similarity matrix, where diagonal of the similarity matrix were 1 (the similarity between two exactly same haplotypes) but the off-diagonals were zeros, was constructed. Similarity matrix for the four haplotypes in **Table 1** was shown in \mathbf{H}_1 .

$$\mathbf{H}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

The next step was transferring the haplotype similarity matrix into individual relatedness matrix. For each pair of individuals, the similarity scores of the four haplotypes harbored by the two individuals were extracted from the haplotype similarity matrices (one of \mathbf{H}_1 , \mathbf{H}_1 , and \mathbf{H}_2) and the relatedness between the two individuals was calculated by summing up the pairwise haplotype alleles similarity scores among the four haplotype alleles and divided by two. Let $\mathbf{W} = \begin{bmatrix} h_{id1Pid2P} & h_{id1Pid2M} \\ h_{id1Mid2P} & h_{id1Mid2M} \end{bmatrix}$ denoted the similarity matrix of the four haplotypes carried by a pair of individuals (id_1 and id_2), where subindexes P and M denoted paternal and maternal haplotype alleles, respectively. The similarity score between id_1 and id_2 was calculated as $S = \frac{h_{id1Pid2P} + h_{id1Pid2M} + h_{id1Mid2P} + h_{id1Mid2M}}{2}$. For example, id_1 in **Table 1** carried *hap1* and *hap4* while id_2 carried *hap2* and *hap1*. The similarity scores of these four haplotypes according to \mathbf{H}_2 were $\mathbf{W} = \begin{bmatrix} 0.17 & 1 \\ 0.71 & 0.22 \end{bmatrix}$ thus the relatedness between id_1

and id_2 was calculated as $G_{HAP,id_1,id_2} = \frac{(0.17+1+0.71+0.22)}{2} = 1.05$. Subsequently, according to \mathbf{H}_2 , the relatedness matrix of individuals shown in **Table 1** could be constructed as \mathbf{G}_{HAP} . Relatedness matrices based on other types of haplotype similarity matrices could be calculated in a similar way.

$$\mathbf{G}_{HAP} = \begin{bmatrix} 1.22 & 1.05 & 0.91 & 0.88 & 0.74 \\ 1.05 & 1.17 & 0.74 & 1.17 & 0.86 \\ 0.91 & 0.74 & 1.36 & 1.01 & 1.18 \\ 0.88 & 1.17 & 1.01 & 2.00 & 1.30 \\ 0.74 & 0.86 & 1.18 & 1.30 & 1.30 \end{bmatrix}$$

The procedures described above constructed the relatedness matrix for one genic haploblock. In practice, relatedness matrices based on the other haploblocks could be built through these procedures and the final genic relatedness matrix was obtained by averaging over the haploblock relatedness matrices. For variance components estimation and genomic prediction, the final relatedness matrix could be easily standardized by dividing the matrix by the maximum of the elements.

Genomic Prediction Models

The statistical model for GP used in this study was

$$\mathbf{y} = \mathbf{1}_n\mu + \mathbf{Z}\mathbf{g} + \mathbf{e}, \quad (1)$$

where \mathbf{y} was a vector of the observations; $\mathbf{1}_n$ was a $n \times 1$ vector with all elements equal to one; μ was the overall mean; \mathbf{Z} was the design matrix allocates observations to genetic values; $\mathbf{g} \sim N(0, \mathbf{K}\sigma_g^2)$ was the genetic values; \mathbf{K} was the relatedness matrices; σ_g^2 was the variance of genetic values; $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$ was the residuals; \mathbf{I} was the identity matrix and σ_e^2 was the residual variance.

We compared the newly proposed approaches to the standard GBLUP (VanRaden, 2008). In GBLUP, the genomic relatedness matrix was calculated as $\mathbf{G} = \frac{\mathbf{M}\mathbf{M}'}{2\sum_{k=1}^m p_k(1-p_k)}$, where \mathbf{M} was the minor allele frequency (MAF) adjusted genotype matrix with elements $(0 - 2p_j)$, $(1 - 2p_j)$, and $(2 - 2p_j)$ representing genotypes AA, AB, and BB, respectively; p_j was the MAF of the j th SNP.

For the genic similarity based models, relatedness matrices were constructed through the procedures described above. These three genic similarity based haplotype models for genomic prediction given gene annotation were denoted as $G_{HAP1|GA}$, $G_{HAP2|GA}$, and $G_{HAP3|GA}$. For comparison, haplotype similarity based relatedness matrices without gene annotation were also calculated. Different from the gene annotation guided approaches, the naïve haplotype models constructed haploblocks for each chromosome starting from the first SNP and the rest steps were the same as genic haplotypes (Meuwissen et al., 2014). The corresponding models without gene annotation were denoted as G_{HAP1} , G_{HAP2} , and G_{HAP3} , respectively.

For all models, variance components were estimated with the *regress* package (Clifford and McCullagh, 2014) in the R platform and genetic values were obtained by solving the mixed model equations.

Assess of Genomic Predictive Ability

Performance of all models were assessed through a 20 times of five-fold cross validation. Variance components were estimated in the training population and genetic values of the test population were predicted via the fitted models. Predictive ability was calculated as the Pearson's correlation between the predicted genetic values and the phenotypic values that pre-adjusted for fixed effects.

Datasets

Rice

Genotypes and phenotypes of the rice breeding population were available from the rice diversity panel (<https://ricediversity.org>) (Begum et al., 2015; Spindel et al., 2015). Briefly, 315 elite rice breeding lines from the International Rice Research Institute (IRRI) irrigated rice breeding program was presented in this rice dataset. Several important traits such as plant height (PH), flower time (FLW), and grain yield (YLD) were tested and recorded in years 2009–2012, including wet and dry seasons each year. Totally, 58,227 SNPs passed the quality control step and were remained for further analysis. The annotations of the latest version of rice genome (*Oryza sativa Japonica Group, Build 4.0*) were downloaded from *Ensemble* via *biomaRt* (Durinck et al., 2005, 2009) R package (Table 2).

Arabidopsis

The Arabidopsis population consisted of 349 natural accessions collected worldwide (Li et al., 2010; Horton et al., 2012; Kooke et al., 2016). Seeds of all accessions were genotyped with 215 K single nucleotide polymorphisms (SNPs; Li et al., 2010; Horton et al., 2012). Three replicates of each accession were cultured and transplanted under the same environmental conditions (Kooke et al., 2016). Lots of developmental traits were measured on all individual plants. Traits used for model comparisons in this study include: leaf area before vernalization (LABv), leaf area after vernalization (LAav), flowering time (FT), petiole to leaf length ratio (PL/LL), petiole length (PL), leaf length (LL), rosette branching (RB), main stem branching (MSB), plant height at 1st silique (PH1S), total plant height (TPH), relative growth rate before vernalization (RGRbv), and relative growth rate after vernalization (RGRav).

Yellow Chicken

The yellow chicken population used in this study was derived from a Chinese indigenous breed and maintained by Wens Nanfang Poultry Breeding Co. Ltd. (Xinxing, P.R. China) (Zhang et al., 2017; Ye et al., 2018). The population consisted of 435

males, which were the 3rd batch of the 25th generation of the population. These birds came from a mixture of full sib and half sib families with the mating of 30 males and 360 females from the 24th generation. After hatching, all birds were maintained in a closed building under controlled environmental conditions and provided with a standard diet till the end of 4 weeks of age. These birds were randomly allocated to three pens for growth performance test from 5 to 13 weeks of age, providing food and water *ad libitum*. After the growth test, all birds were slaughtered at the age of 91 days. Seventeen traits including average daily gain (ADG), average daily feed intake (ADFI), residual feed intake (RFI), and intestine length (IL) were used for model validation in this study. All individuals were genotyped with the commercially available 600 K Affymetrix Axion HD genotyping array using DNA extracted from blood samples. The phenotypes were pre-adjusted for the fixed pen effect via the following statistical model:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \mathbf{e},$$

where \mathbf{y} was a vector the raw phenotypes; \mathbf{X} and \mathbf{Z} were design matrices; \mathbf{b} was a vector of the fixed pen effects; $\mathbf{u} \sim N(0, \mathbf{G}\sigma_u^2)$ was the vector of genetic values; \mathbf{G} was the SNP derived relatedness matrix (VanRaden, 2008); σ_u^2 was the additive genetic variance; $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$ was the vector of residuals; σ_e^2 was the residual variance and \mathbf{I} was the identity matrix. The adjusted phenotypes $\bar{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}}$ were used as model response in the genomic prediction models.

RESULTS

Predictive Ability in the Rice Population

Predictive ability of all models in the rice breeding population was shown in Table 3 and Figure S1. Overall, the gene annotation based haplotype models ($\sim|GA$ models) outperformed *GBLUP* and the naïve haplotype models to some extent. Among the three gene annotation based haplotype models, $G_{HAPI|GA}$, where an identity matrix was used to measure similarity between pairs of haplotype alleles, performed best in respect of predictive ability. For plant height, $G_{HAPI|GA}$ showed the highest predictive ability. Compared to *GBLUP*, 4.73 and 6.43% extra accuracy were obtained by incorporating gene annotation in a haplotype model for dry season (DS_PH) and wet season (WS_PH), respectively; $G_{HAPI|GA}$ improved 2.21% (DS_PH) and 3.43% (WS_PH) of the predictive ability compared to the naïve haplotype model G_{HAPI} . For flowering time, $G_{HAPI|GA}$ was 5.62 and 7.07% higher than *GBLUP* in respect of predictive ability in dry season and wet season, respectively; $G_{HAPI|GA}$ outperformed G_{HAPI} by 1.62 and

TABLE 2 | Datasets description.

| Datasets | # of observations | # of markers | Reference genome | # of mapped SNPs | # of represented genes | # of haplotypes |
|-------------|-------------------|--------------|---|------------------|------------------------|-----------------|
| Rice | 315 | 58,227 | <i>Oryza sativa</i> Japonica Group (Build 4.0) | 44,831 | 22,509 | 25,453 |
| Arabidopsis | 349 | 208,481 | <i>Arabidopsis thaliana</i> (assembly TAIR10.1) | 193,646 | 27,169 | 167,837 |
| Chicken | 435 | 408,715 | <i>Gallus gallus</i> (assembly GGA 5) | 233,417 | 17,686 | 45,470 |

Denoted "the number."

TABLE 3 | Pearson's correlation between observed and predicted phenotypes in the rice breeding population (Mean \pm SE).

| Traits | GBLUP ^a | G _{HAPI} ^b | G _{HAPI} GA ^c | G _{HAP1} ^b | G _{HAP1} GA ^c | G _{HAP2} ^b | G _{HAP2} GA ^c |
|--------|--------------------|--------------------------------|-------------------------------------|--------------------------------|-------------------------------------|--------------------------------|------------------------------------|
| DS_PH | 0.486 \pm 0.007 | 0.498 \pm 0.007 | 0.509 \pm 0.007 | 0.493 \pm 0.007 | 0.501 \pm 0.007 | 0.498 \pm 0.007 | 0.503 \pm 0.007 |
| DS_FLW | 0.534 \pm 0.005 | 0.555 \pm 0.005 | 0.564 \pm 0.005 | 0.530 \pm 0.005 | 0.552 \pm 0.005 | 0.540 \pm 0.005 | 0.553 \pm 0.005 |
| DS_YLD | 0.289 \pm 0.006 | 0.285 \pm 0.006 | 0.313 \pm 0.006 | 0.286 \pm 0.006 | 0.312 \pm 0.006 | 0.286 \pm 0.006 | 0.311 \pm 0.006 |
| WS_PH | 0.482 \pm 0.006 | 0.496 \pm 0.005 | 0.513 \pm 0.005 | 0.489 \pm 0.006 | 0.507 \pm 0.005 | 0.492 \pm 0.006 | 0.509 \pm 0.005 |
| WS_FLW | 0.467 \pm 0.007 | 0.487 \pm 0.006 | 0.500 \pm 0.006 | 0.465 \pm 0.006 | 0.491 \pm 0.006 | 0.474 \pm 0.006 | 0.492 \pm 0.006 |
| WS_YLD | 0.258 \pm 0.007 | 0.242 \pm 0.007 | 0.268 \pm 0.008 | 0.264 \pm 0.007 | 0.282 \pm 0.008 | 0.256 \pm 0.007 | 0.280 \pm 0.008 |

For each trait (row), the values in boldface indicate the best prediction among all models. DS, dry season; WS, wet season; PH, plant height; FLW, flower time; YLD, grain yield.

^aGenomic best linear unbiased prediction (VanRaden, 2008).

^bHaplotype similarity based models without gene annotation. HAPI, HAP1, and Hap2 are differ on the way of evaluating haplotype similarity.

^c~|GA denoted gene annotation guided GP models.

2.67% in DS_FLW and WS_FLW, respectively. For grain yield, G_{HAPI}|GA showed the highest predictive ability in DS_YLD, which was 8.30 and 9.82% higher than GBLUP and G_{HAPI}, respectively; G_{HAPI}|GA showed the best predictive ability in WS_YLD, which was 9.30% and 16.53% higher than GBLUP and G_{HAPI}, respectively.

Predictive Ability in the Arabidopsis Population

Table 4 and Figure S2 showed the predictive ability in the Arabidopsis population. Overall, gene annotation based haplotype models outperformed GBLUP and the naïve haplotype models in 8 out of 12 traits. For Laav, PH1S, MSB, FT, and RGRbv, G_{HAPI}|GA showed the best performance in respect of predictive ability and outperformed GBLUP by 5.47, 13.61, 5.59, 3.58, and 26.67%, respectively. For LL, PL, and RGRav, G_{HAPI}|GA showed the best performance and outperformed GBLUP by 0.56%, 1.98% and 9.78%, respectively. However, G_{HAPI}, G_{HAP1}, and G_{HAP2}, in which gene annotation information was not integrated, outperformed GBLUP and the gene annotation based haplotype models (~|GA) for the traits RB, PL/LL, and TPH.

Predictive Ability in the Yellow Chicken Population

Table 5 and Figure S3 showed the predictive ability in the yellow chicken population. The haplotype models benefit from gene annotation information in six (MTW, MTMW, RFI, EW, DW, and BW45) out of 17 traits, where gene annotation models outperformed GBLUP and naïve haplotype models by 0.40~3.43%. For ADG, ADFI, EWG, BMW, AFW, and IL, GBLUP showed the best performance, while haplotype models with or without gene annotation did not show any extra gain in respect of predictive ability. For RFI and FCR, G_{HAPI} was slightly better than GBLUP.

DISCUSSION

In this study, SNPs were mapped to protein coding genes according to the physical positions and used for haplotype construction. Different from our previous study (Gao et al., 2017), in which genic region haplotypes were encoded in both a numerical and a categorical strategy, here we constructed

individual similarity matrices from the haplotype allele similarity matrices via strategies described by Hickey et al. (2013). Three strategies were utilized to calculate similarity scores between haplotype alleles. Individual similarity matrices were constructed by averaging the haplotype similarity among all genes or genome regions and used in the genetic evaluations.

Generally, the gene annotation based haplotype models proposed in this study potentially improved the genomic predictive ability. In the three datasets of rice, Arabidopsis, and yellow chicken, gene annotation based models improved the predictive ability in several traits, especially traits in the rice breeding population (Table 3 and Figure S1) and the Arabidopsis population (Table 4 and Figure S2), compared to GBLUP model. Results in the rice dataset showed that incorporating gene annotation in a haplotype model could improve the predictive ability. However, the extent of improvement was slightly lower compared to the categorical models in Gao et al. (2017). The phenomenon could be explained by two possible reasons. Firstly, non-additive effects played important roles in controlling the plant traits (Shen et al., 2014). Dominance and epistasis were additionally considered in the previous gene annotation based categorical models (Gao et al., 2017). The impact of non-additive effects on predictive ability could also be seen when comparing the performance of haplotype allele dosage models with categorical epistasis models in Gao et al. (2017). Secondly, the haplotype allele similarity scores could more or less reflect the identical by descent (IBD) between SNP alleles and thus better in measuring relatedness between pairs of individuals (de Roos et al., 2011). However, the advantages on similarity measuring were not always transferred into the predictive ability (Hickey et al., 2013).

However, integrating gene annotation just slightly improved the predictive ability of several traits and did not show any improvement in the rest in the yellow chicken population (Table 5 and Figure S3). The possible reasons were the frequent recombination in the chicken genome (Fulton et al., 2016) and the underlying trait genetic architecture. Generally speaking, haplotype models were more powerful on reflecting real relatedness between individuals. However, the advantages of haplotype derived relatedness matrices could be expected only when haplotypes were better in tracing the underlying recombination events than SNPs. Previous studies have found

TABLE 4 | Pearson's correlation between observed and predicted phenotypes in the Arabidopsis population (Mean \pm SE).

| Traits | GBLUP ^a | G _{HAP1} ^b | G _{HAP1 GA} ^c | G _{HAP1} ^b | G _{HAP1 GA} ^c | G _{HAP2} ^b | G _{HAP2 GA} ^c |
|--------|--------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-----------------------------------|
| Labv | 0.163 \pm 0.009 | 0.161 \pm 0.009 | 0.170 \pm 0.009 | 0.164 \pm 0.009 | 0.176 \pm 0.009 | 0.166 \pm 0.009 | 0.174 \pm 0.009 |
| Laav | 0.201 \pm 0.006 | 0.205 \pm 0.006 | 0.212 \pm 0.005 | 0.200 \pm 0.006 | 0.209 \pm 0.006 | 0.201 \pm 0.006 | 0.208 \pm 0.006 |
| PH1S | 0.191 \pm 0.005 | 0.196 \pm 0.005 | 0.217 \pm 0.005 | 0.190 \pm 0.005 | 0.213 \pm 0.005 | 0.191 \pm 0.005 | 0.211 \pm 0.005 |
| TPH | 0.185 \pm 0.007 | 0.183 \pm 0.007 | 0.175 \pm 0.007 | 0.186 \pm 0.007 | 0.181 \pm 0.007 | 0.185 \pm 0.007 | 0.179 \pm 0.007 |
| MSB | 0.340 \pm 0.004 | 0.346 \pm 0.004 | 0.359 \pm 0.004 | 0.337 \pm 0.004 | 0.346 \pm 0.004 | 0.337 \pm 0.004 | 0.348 \pm 0.004 |
| RB | 0.281 \pm 0.006 | 0.289 \pm 0.007 | 0.283 \pm 0.007 | 0.281 \pm 0.007 | 0.277 \pm 0.006 | 0.282 \pm 0.006 | 0.276 \pm 0.006 |
| LL | 0.356 \pm 0.006 | 0.355 \pm 0.005 | 0.353 \pm 0.005 | 0.356 \pm 0.006 | 0.358 \pm 0.006 | 0.358 \pm 0.006 | 0.357 \pm 0.005 |
| PL | 0.303 \pm 0.006 | 0.301 \pm 0.006 | 0.301 \pm 0.005 | 0.305 \pm 0.006 | 0.309 \pm 0.006 | 0.306 \pm 0.006 | 0.307 \pm 0.006 |
| PL/LL | 0.255 \pm 0.009 | 0.249 \pm 0.009 | 0.237 \pm 0.008 | 0.258 \pm 0.010 | 0.247 \pm 0.008 | 0.257 \pm 0.009 | 0.245 \pm 0.008 |
| FT | 0.643 \pm 0.003 | 0.653 \pm 0.003 | 0.666 \pm 0.003 | 0.642 \pm 0.003 | 0.658 \pm 0.003 | 0.644 \pm 0.003 | 0.660 \pm 0.003 |
| RGRbv | 0.045 \pm 0.007 | 0.050 \pm 0.007 | 0.057 \pm 0.007 | 0.042 \pm 0.007 | 0.054 \pm 0.008 | 0.042 \pm 0.007 | 0.054 \pm 0.008 |
| RGRav | 0.184 \pm 0.006 | 0.179 \pm 0.006 | 0.194 \pm 0.006 | 0.184 \pm 0.006 | 0.202 \pm 0.006 | 0.183 \pm 0.006 | 0.199 \pm 0.006 |

For each trait (row), the values in boldface indicate the best prediction among all models. Labv, leaf area before vernalization; Laav, leaf area after vernalization; FT, flowering time; PL/LL, petiole to leaf length ratio; PL, petiole length; LL, leaf length; RB, rosette branching; MSB, main stem branching; PH1S, plant height at 1st silique; TPH, total plant height; RGRbv, relative growth rate before vernalization; RGRav, relative growth rate after vernalization.

^aGenomic best linear unbiased prediction (VanRaden, 2008).

^bHaplotype similarity based models without gene annotation. HAPI, HAP1, and Hap2 are differ on the way of evaluating haplotype similarity.

^c~|GA denoted gene annotation guided GP models.

TABLE 5 | Pearson's correlation between observed and predicted phenotypes in the yellow chicken population (Mean \pm SE).

| Traits | GBLUP ^a | G _{HAP1} ^b | G _{HAP1 GA} ^c | G _{HAP1} ^b | G _{HAP1 GA} ^c | G _{HAP2} ^b | G _{HAP2 GA} ^c |
|--------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|--------------------------------|-------------------------------------|
| ADG | 0.351 \pm 0.005 | 0.344 \pm 0.005 | 0.342 \pm 0.004 | 0.345 \pm 0.005 | 0.345 \pm 0.004 | 0.345 \pm 0.005 | 0.345 \pm 0.004 |
| ADFI | 0.440 \pm 0.004 | 0.437 \pm 0.004 | 0.438 \pm 0.004 | 0.436 \pm 0.004 | 0.439 \pm 0.004 | 0.437 \pm 0.004 | 0.440 \pm 0.004 |
| MTW | 0.322 \pm 0.005 | 0.315 \pm 0.004 | 0.328 \pm 0.004 | 0.314 \pm 0.005 | 0.325 \pm 0.004 | 0.316 \pm 0.005 | 0.326 \pm 0.004 |
| MTMW | 0.322 \pm 0.005 | 0.315 \pm 0.004 | 0.328 \pm 0.004 | 0.314 \pm 0.005 | 0.325 \pm 0.004 | 0.316 \pm 0.005 | 0.327 \pm 0.004 |
| RFI | 0.464 \pm 0.005 | 0.468 \pm 0.005 | 0.468 \pm 0.005 | 0.465 \pm 0.005 | 0.466 \pm 0.005 | 0.467 \pm 0.005 | 0.467 \pm 0.005 |
| FCR | 0.288 \pm 0.004 | 0.289 \pm 0.004 | 0.274 \pm 0.004 | 0.286 \pm 0.004 | 0.271 \pm 0.004 | 0.288 \pm 0.004 | 0.273 \pm 0.004 |
| EWG | 0.257 \pm 0.009 | 0.253 \pm 0.009 | 0.256 \pm 0.009 | 0.253 \pm 0.009 | 0.256 \pm 0.008 | 0.254 \pm 0.009 | 0.256 \pm 0.009 |
| EW | 0.253 \pm 0.009 | 0.249 \pm 0.010 | 0.253 \pm 0.009 | 0.250 \pm 0.010 | 0.254 \pm 0.009 | 0.250 \pm 0.010 | 0.254 \pm 0.009 |
| BMW | 0.144 \pm 0.011 | 0.142 \pm 0.011 | 0.138 \pm 0.011 | 0.144 \pm 0.011 | 0.142 \pm 0.011 | 0.143 \pm 0.011 | 0.141 \pm 0.011 |
| BMP | 0.128 \pm 0.011 | 0.128 \pm 0.011 | 0.123 \pm 0.011 | 0.130 \pm 0.011 | 0.128 \pm 0.011 | 0.129 \pm 0.011 | 0.126 \pm 0.011 |
| DW | 0.175 \pm 0.010 | 0.172 \pm 0.010 | 0.176 \pm 0.010 | 0.175 \pm 0.010 | 0.181 \pm 0.009 | 0.174 \pm 0.010 | 0.179 \pm 0.010 |
| DP | 0.128 \pm 0.011 | 0.128 \pm 0.011 | 0.123 \pm 0.011 | 0.130 \pm 0.011 | 0.128 \pm 0.011 | 0.129 \pm 0.011 | 0.126 \pm 0.011 |
| AFW | 0.114 \pm 0.009 | 0.108 \pm 0.009 | 0.104 \pm 0.009 | 0.112 \pm 0.009 | 0.110 \pm 0.009 | 0.111 \pm 0.009 | 0.108 \pm 0.009 |
| AFP | 0.128 \pm 0.011 | 0.128 \pm 0.011 | 0.123 \pm 0.011 | 0.130 \pm 0.011 | 0.128 \pm 0.011 | 0.129 \pm 0.011 | 0.126 \pm 0.011 |
| GW | 0.067 \pm 0.011 | 0.070 \pm 0.010 | 0.066 \pm 0.011 | 0.071 \pm 0.010 | 0.068 \pm 0.011 | 0.070 \pm 0.011 | 0.067 \pm 0.011 |
| IL | 0.045 \pm 0.005 | 0.041 \pm 0.005 | 0.037 \pm 0.005 | 0.043 \pm 0.005 | 0.040 \pm 0.005 | 0.043 \pm 0.005 | 0.039 \pm 0.005 |
| BW45 | 0.307 \pm 0.005 | 0.306 \pm 0.005 | 0.309 \pm 0.005 | 0.303 \pm 0.005 | 0.302 \pm 0.005 | 0.304 \pm 0.005 | 0.304 \pm 0.005 |

For each trait (row), the values in boldface indicate the best prediction among all models. ADG, Average daily gain; ADFI, Average daily feed intake; MTW, Mid-term body weight; MTMW, Mid-term metabolic body weight; RFI, Residual feed intake; FCR, Feed conversion rate; EWG, Eviscerated weight with giblet; EW, Eviscerated weight; BMW, Breast muscle weight; BMP, Breast muscle percentage; DW, Drumstick weight; DP, Drumstick percentage; AFW, Abdominal fat weight; AFP, Abdominal fat percentage; GW, Gizzard weight; IL, intestine length; BW45, body weight at 45 day.

^aGenomic best linear unbiased prediction (VanRaden, 2008).

^bHaplotype similarity based models without gene annotation. HAPI, HAP1, and Hap2 are differ on the way of evaluating haplotype similarity.

^c~|GA denoted gene annotation guided GP models.

extensive diversity and large number of recombination hotspots in the chicken genome (Fulton et al., 2016), which shorten the real haplotype blocks and thus linkage disequilibrium based approaches were more suitable for haplotype blocks constructing. In this study, instead of considering linkage disequilibrium, we implemented a strategy similar to Meuwissen et al. (2014),

where maximum number of haplotype alleles was used as threshold when adding SNPs to haplotypes, for haploblock constructing. This approach might not be suitable for the species that extensive diversity and abundant recombination existed in the genome. Therefore, linkage disequilibrium based haploblock construction methods (Cuyabano et al., 2015; Da,

2015) should be suggested for such species. Nevertheless, the main focus of this study was to provide methods of building genic similarity relationship matrices, though the haplotype could be defined through various rules. Even the setting of threshold of the number of haplotype alleles harbored in each haploblock was relatively arbitrary, it was an easy way to build haplotypes and good at controlling the number of variables within each haploblock. Actually, LD information was also reflected indirectly by restricting the maximum of haplotype alleles in certain haploblock, since lower LD among consecutive SNPs would increase the number of haplotype alleles rapidly when adding more SNPs to the haploblocks. Moreover, to our knowledge, the LD based haplotype construction method might have problems on inadequate accurate estimations of LD level in small populations and difficulty in selecting LD threshold for combining consecutive SNPs into haplotype.

In this study, the relatedness matrices used for genetic evaluation were constructed by averaging the relatedness based on individual genes, which meant that weights were assigned equally among genes. The underlying assumption of this approach was that all genes contributed equally to the relatedness matrices and thus to the traits. However, abundant accumulative biological knowledge had shown that gene effects were different among traits. Moreover, previous studies had found that genomic prediction models could be improved when genetic architecture was considered by assigning different weights to SNPs (Zhang et al., 2010; Ober et al., 2012; Gao et al., 2015). Therefore, similar approaches to construct trait specific relatedness matrices by weighting genes differently (Zhang et al., 2010; Ober et al., 2012; Gao et al., 2015) in the paradigm of genic similarity genomic prediction models are worth trying in the future.

Overall, we proposed a new strategy to construct relatedness matrices on the gene level by transferring the genic haplotype similarity scores into individual similarity matrices. New explanatory variables on the gene level were derived from phased SNPs and through which the prediction model was moved one step further from SNPs to biologically functional units. The genic similarity matrices based model showed benefit in respect of predictive ability for many traits in the studied populations. However, predictive ability was not improved in some traits, especially in the yellow chicken population, which indicated that the newly proposed approach

still had rooms for improvement to adapt different traits or populations. The uniform weight assigned among genes when constructing the relatedness matrices and the insensitivity to genome recombination rate (the strategy for genic haplotype construction) could be the two major limitations of the new approach. Nevertheless, the idea of constructing relatedness matrices on the biologically functional units potentially improved predictive ability.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of Animal Care Committee of South China Agriculture University (Guangzhou, People's Republic of China). The protocol was approved by the Animal Care Committee of South China Agriculture University. Animals involved in this study were humanely sacrificed as necessary to ameliorate their suffering.

AUTHOR CONTRIBUTIONS

NG and ZZ conceived this study, performed the model validations, and wrote the manuscript. JT and SH helped in the model validations and manuscript. SY, HZ, and XY helped in the manuscript writing. XZ originally derived the yellow chicken data and helped in the analyses. JL stimulated the idea of the paper and helped in the manuscript.

ACKNOWLEDGMENTS

This study is funded by National Natural Science Foundation of China (31772556), the earmarked fund for China Agriculture Research System (CARS-35, CARS-41), and Guangdong Sailing Program (2014YT02H042). We would like to thank professor Joost J. B. Keurentjes for providing the Arabidopsis data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2018.00364/full#supplementary-material>

REFERENCES

- Abdollahi-Arpanahi, R., Morota, G., and Peñagaricano, F. (2017). Predicting bull fertility using genomic data and biological information. *J. Dairy Sci.* 100, 9656–9666. doi: 10.3168/jds.2017-13288
- Abdollahi-Arpanahi, R., Morota, G., Valente, B. D., Kranis, A., Rosa, G. J., and Gianola, D. (2016). Differential contribution of genomic regions to marked genetic variation and prediction of quantitative traits in broiler chickens. *Genet. Sel. Evol.* 48:10. doi: 10.1186/s12711-016-0187-z
- Begum, H., Spindel, J. E., Lalusin, A., Borromeo, T., Gregorio, G., Hernandez, J., et al. (2015). Genome-wide association mapping for yield and other agronomic traits in an elite breeding population of tropical rice (*Oryza sativa*). *PLoS ONE* 10:e0119873. doi: 10.1371/journal.pone.0119873
- Calus, M. P., Meuwissen, T. H., de Roos, A. P., and Veerkamp, R. F. (2008). Accuracy of genomic selection using different methods to define haplotypes. *Genetics* 178, 553–561. doi: 10.1534/genetics.107.080838
- Clifford, D., and McCullagh, P. (2014). *The Regress package R Package Version 1*. 3–14. Available online at: <https://cran.r-project.org/web/packages/regress/citation.html>
- Cuyabano, B. C., Su, G., and Lund, M. S. (2015). Selection of haplotype variables from a high-density marker map for genomic prediction. *Genet. Sel. Evol.* 47:61. doi: 10.1186/s12711-015-0143-3
- Da, Y. (2015). Multi-allele haplotype model based on genetic partition for genomic prediction and variance component estimation using SNP markers. *BMC Genet.* 16:144. doi: 10.1186/s12863-015-0301-1
- de Los Campos, G., Vazquez, A. I., Fernando, R., Klimentidis, Y. C., and Sorensen, D. (2013). Prediction of complex human traits using

- the genomic best linear unbiased predictor. *PLoS Genet.* 9:e1003608. doi: 10.1371/journal.pgen.1003608
- de Roos, A. P., Schrooten, C., and Druet, T. (2011). Genomic breeding value estimation using genetic markers, inferred ancestral haplotypes, and the genomic relationship matrix. *J. Dairy Sci.* 94, 4708–4714. doi: 10.3168/jds.2010-3905
- Do, D. N., Janss, L. L., 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
- Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., et al. (2005). BioMart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 21, 3439–3440. doi: 10.1093/bioinformatics/bti525
- 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, S. M., Sørensen, I. F., Sarup, P., Mackay, T. F. C., and Sørensen, 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
- Fulton, J. E., McCarron, A. M., Lund, A. R., Pinegar, K. N., Wolc, A., Chazara, O., et al. (2016). A high-density SNP panel reveals extensive diversity, frequent recombination and multiple recombination hotspots within the chicken major histocompatibility complex B region between BG2 and CD1A1. *Genet. Sel. Evol.* 48:1. doi: 10.1186/s12711-015-0181-x
- 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 Genet.* 16:120. doi: 10.1186/s12863-015-0278-9
- Gao, N., Martini, J. W. R., Zhang, Z., Yuan, X., Zhang, H., Simianer, H., et al. (2017). Incorporating gene annotation into genomic prediction of complex phenotypes. *Genetics* 207, 489–501. doi: 10.1534/genetics.117.300198
- Gianola, D. (2013). Priors in whole-genome regression: the Bayesian alphabet returns. *Genetics* 194, 573–596. doi: 10.1534/genetics.113.151753
- Gianola, D., Fernando, R. L., and Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics* 173, 1761–1776. doi: 10.1534/genetics.105.049510
- Gianola, D., and van Kaam, J. B. (2008). Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics* 178, 2289–2303. doi: 10.1534/genetics.107.084285
- Habier, D., Fernando, R. L., Kizilkaya, K., and Garrick, D. J. (2011). Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics* 12:186. doi: 10.1186/1471-2105-12-186
- Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* 31:423. doi: 10.2307/2529430
- Hickey, J. M., Kinghorn, B. P., Tier, B., Clark, S. A., van der Werf, J. H., and Gorjanc, G. (2013). Genomic evaluations using similarity between haplotypes. *J. Anim. Breed. Genet.* 130, 259–269. doi: 10.1111/jbg.12020
- Horton, M. W., Hancock, A. M., Huang, Y. S., Toomajian, C., Atwell, S., Auton, A., et al. (2012). Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nat. Genet.* 44, 212–216. doi: 10.1038/ng.1042
- Hu, Z. L., Park, C. A., and Reacy, J. M. (2016). Developmental progress and current status of the Animal QTLdb. *Nucleic Acids Res.* 44, D827–D833. doi: 10.1093/nar/gkv1233
- Kooke, R., Kruijer, W., Bours, R., Becker, F., Kuhn, A., van de Geest, H., et al. (2016). Genome-Wide association mapping and genomic prediction elucidate the genetic architecture of morphological traits in *Arabidopsis*. *Plant Physiol.* 170, 2187–2203. doi: 10.1104/pp.15.00997
- Li, Y., Huang, Y., Bergelson, J., Nordborg, M., and Borevitz, J. O. (2010). Association mapping of local climate-sensitive quantitative trait loci in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U.S.A.* 107, 21199–21204. doi: 10.1073/pnas.1007431107
- MacLeod, I. M., Bowman, P. J., Vander Jagt, C. J., 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
- 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.
- Meuwissen, T. H., Odegard, J., Andersen-Ranberg, I., and Grindflek, E. (2014). On the distance of genetic relationships and the accuracy of genomic prediction in pig breeding. *Genet. Sel. Evol.* 46:49. doi: 10.1186/1297-9686-46-49
- 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
- Ober, U., Ayroles, J. F., Stone, E. A., Richards, S., Zhu, D., Gibbs, R. A., et al. (2012). Using whole-genome sequence data to predict quantitative trait phenotypes in *Drosophila melanogaster*. *PLoS Genet.* 8:e1002685. doi: 10.1371/journal.pgen.1002685
- R Development Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
- Ramstein, G. P., Evans, J., Kappler, S. M., Mitchell, R. B., Vogel, K. P., Buell, C. R., et al. (2016). Accuracy of genomic prediction in switchgrass (*Panicum virgatum* L.) improved by accounting for linkage disequilibrium. *G3* 6, 1049–1062. doi: 10.1534/g3.115.024950
- Shen, G., Zhan, W., Chen, H., and Xing, Y. (2014). Dominance and epistasis are the main contributors to heterosis for plant height in rice. *Plant Sci.* 215–216, 11–18. doi: 10.1016/j.plantsci.2013.10.004
- Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redoña, E., et al. (2015). Genomic selection and association mapping in rice (*Oryza sativa*): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PLoS Genet.* 11:e1004982. doi: 10.1371/journal.pgen.1004982
- VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi: 10.3168/jds.2007-0980
- Ye, S., Yuan, X., Lin, X., Gao, N., Luo, Y., Chen, Z., et al. (2018). Imputation from SNP chip to sequence: a case study in a Chinese indigenous chicken population. *J. Anim. Sci. Biotechnol.* 9:30. doi: 10.1186/s40104-018-0241-5
- Zhang, Z., Liu, J., Ding, X., Bijma, P., de Koning, D. J., and Zhang, Q. (2010). Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. *PLoS ONE* 5:e12648. doi: 10.1371/journal.pone.0012648
- 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
- Zhang, Z., Xu, Z. Q., Luo, Y. Y., Zhang, H. B., Gao, N., He, J. L., et al. (2017). Whole genomic prediction of growth and carcass traits in a chinese quality chicken population. *J. Anim. Sci.* 95, 72–80. doi: 10.2527/jas2016.0823

Conflict of Interest Statement: 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.

Copyright © 2018 Gao, Teng, Ye, Yuan, Huang, Zhang, Zhang, Li and Zhang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.