Multi-omics-data-assisted genomic feature preselection improves the accuracy of genomic prediction

Shaopan Ye
South China Agricultural University https://orcid.org/0000-0001-6333-206X
Jiaqi Li
South China Agricultural University
Zhe Zhang (zhezhang@scau.edu.cn)
South China Agricultural University

Research Article

Keywords: multi-omics data, SNP preselection, genomic prediction, accuracy, Drosophila melanogaster.

Posted Date: May 5th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-26522/v1

Multi-omics-data-assisted genomic feature preselection improves the accuracy of genomic prediction

Shaopan Ye[1], Jiaqi Li [1], and Zhe Zhang[1*]

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

*Corresponding author: Zhe Zhang

Email addresses:

SPY: yy623689080@126.com

JQL: jqli@cau.edu.cn

ZZ: zhezhang@scau.edu.cn

Abstract

Background: Presently, multi-omics data (e.g., genomics, transcriptomics, proteomics, and metabolomics) are available for genomic prediction. Omics data not only offer new data layers for genomic prediction but also provide a bridge between organismal phenotypes and genome variation that cannot be readily captured at the genome sequence level. Therefore, using multi-omics data to select feature markers is a feasible strategy to improve the accuracy of genomic prediction. In this study, simultaneously using whole-genome sequencing (WGS) and gene expression level data, four strategies for single-nucleotide polymorphism (SNP) preselection were investigated for genomic predictions in the Drosophila Genetic Reference Panel.

Results: Using genomic best linear unbiased prediction (GBLUP) with complete WGS data, the prediction accuracy values were 0.208±0.020 (0.181±0.022) for the startle response and 0.272±0.017 (0.307±0.015) for starvation resistance in the female (male) lines. Compared with GBLUP using complete WGS data, both GBLUP and the genomic feature BLUP (GFBLUP) did not improve the prediction accuracy using SNPs preselected from the complete WGS data based on the results of genome-wide association studies (GWASs) or transcriptome-wide association studies (TWASs). Furthermore, by using SNPs preselected from the WGS data based on the results of the expression quantitative trait locus (eQTL) mapping of all genes, only the startle response had greater accuracy than GBLUP with the complete WGS data. The best accuracy values in the female and male lines were 0.243±0.020 and 0.220±0.022, respectively. Importantly, by using SNPs preselected based on the results of the eQTL

mapping of significant genes from TWAS, both GBLUP and GFBLUP resulted in a greater accuracy and smaller bias of genomic prediction. For the startle response, the best accuracy values were 0.258±0.019 (0.237±0.019) for GBLUP and 0.265±0.018 (0.245±0.020) for GFBLUP in the female (male) lines. For starvation resistance, the best accuracy values were 0.437±0.015 (0.427±0.015) for GBLUP and 0.419±0.016 (0.390±0.014) for GFBLUP in female (male) lines. Compared to the GBLUP with complete WGS data, the best accuracy values represented increases of 60.66% and 39.09% for the startle response and 27.40% and 35.36% for starvation resistance in the female and male lines, respectively.

Conclusions: Overall, multi-omics data can assist genomic feature preselection and improve the performance of genomic prediction. The new knowledge gained from this study will enrich the use of multi-omics in genomic prediction.

Keywords: multi-omics data, SNP preselection, genomic prediction, accuracy, Drosophila melanogaster.

Introduction

Genomic prediction, also known as genomic selection (GS), was initially proposed in (Meuwissen et al., 2001) and is a statistical method to predict the yet-to-be observed phenotypes or unobserved genetic values of complex traits based on genomic data. The assumption of this method is that all quantitative trait loci (QTLs) are in linkage disequilibrium (LD) with at least one marker in the whole genome. GS is famous for shortening the generation intervals and increasing the reliability of predicted breeding values, especial for dairy cattle breeding (Garcia-Ruiz et al., 2016). Presently, genomic prediction is widely used in animal and plant breeding and polygenic disease risk prediction.

Over the past decade, the implementation of GS was mainly based on single-nucleotide polymorphism (SNP) chip data. With the cost of sequencing dropping rapidly, it became possible to perform genomic predictions with whole-genome sequencing (WGS) data. Compared with SNP chip data, WGS data are expected to improve the accuracy of genomic predictions by increasing the level of LD between the SNPs and a QTL, even including causal mutations. Simulation studies confirmed the hypothesis that WGS data would improve the accuracy of genomic prediction in a single population (Meuwissen and Goddard, 2010) or multi-populations (Iheshiulor et al., 2016). However, using real WGS data, a higher accuracy of genomic prediction was not achieved for Drosophila (Ober et al., 2012), and similar results were found for livestock using imputed WGS data (van Binsbergen et al., 2015; Zhang et al., 2018; Ye et al., 2019). Possibly, large amounts of markers are both non-causal makers and not in LD with the causal loci.

Moreover, our previous study indicated that the LD pruning of imputed WGS data could improve prediction accuracy (Ye et al., 2019). Therefore, pre-selected potential causal markers or QTLs from WGS are necessary for improving the accuracy of genomic prediction (Raymond et al., 2018). Thus, many preselection variant strategies were used to improve the power of genomic prediction based on the following methods: genome-wide association study (GWAS) (Zhang et al., 2014; Veerkamp et al., 2016; Song et al., 2019; Ye et al., 2019), Bayesian procedures (Kemper et al., 2015), genome-wide signatures of selection (Ye et al., 2020), Animal QTLdb (Song et al., 2019), gene annotation (Heidaritabar et al., 2016; Gao et al., 2017), and gene ontology categories (Edwards et al., 2016; Abdollahiarpanahi et al., 2017). These methods mainly depend on the direct link between phenotype and DNA variants or some prior genome annotation information. However, the genetic links between phenotype and genome variants are too complex to determine directly at the genome sequencing level.

Presently, it has become possible to obtain multi-omics data (e.g., genomic, transcriptomics, proteomics, and metabolomics) for genomic predictions. This makes it possible to uncover genotype–phenotype relationships using different types of data. Related studies were reported using omics data to perform genomic prediction for complex traits in humans (Vazquez et al., 2016; Dimitrakopoulos et al., 2017), plants (Xu et al., 2017; Azodi et al., 2019; Hu et al., 2019; Wang et al., 2019), and model animals (Li et al., 2019; Morgante et al., 2019). Most of these studies focused on integrating multiple omics data into a prediction model to improve prediction accuracy (Guo et al., 2016; Xu et al., 2017; Li et al., 2019; Morgante et al., 2019). However, multi-

omics data not only offer new data layers for genomic prediction but also provide a bridge between organismal phenotype and genome variation that cannot be readily captured at the genome sequence level (Azodi et al., 2019). Therefore, using omics data to select feature makers is a feasible strategy to improve the accuracy of genomic prediction.

In this study, using WGS and gene expression level data, different strategies of SNP preselection were investigated for genomic predictions in the Drosophila genetic reference panel (DGRP). Our results provide useful knowledge about preselected genomic features based on multi-omics data and thus improve the power of genomic predictions for complex traits.

Materials and methods

The genomic, transcriptomic, and phenotypic data of DGRP lines

The DGRP is a living library of common polymorphisms affecting complex traits, as well as a community resource for the whole genome association mapping of quantitative trait loci (Mackay et al., 2012; Huang et al., 2014). The DGRP has 205 inbred lines derived from 20 generations of full-sib mating from isofemale lines collected at the Farmer's Market in Raleigh, NC, USA. These 205 lines were subjected to whole genome sequencing using Illumina and 454 sequencing. After variant calling, a total of 4,672,297 SNPs were found around the chromosome arm (X, 2L, 2R, 3L, 3R, 4) (Mackay et al., 2012). The gene expression level of 200 DGRP lines (as the $\log_2$-transformed fragments per kilobase of transcript per million fragments mapped, FPKM) for 15,732 genes in females and 20,375 genes in males were obtained by Everett et al

(2020). Furthermore, two traits (startle response and starvation resistance) were selected as model traits. Finally, totals of 198 and 199 lines for starvation resistance and startle response, respectively, were used for further genomic prediction due to allowing the measurement of phenotypes and expression levels simultaneously. The quality control of the WGS data was conducted using PLINK (Purcell et al., 2007) with the criteria of SNP call rate ≥ 95%, individual call rate ≥ 97%, MAF ≥ 5%, and the Hardy–Weinberg equilibrium P-value ≥ 1.0e-6. The missing genotypes were imputed by Beagle 4.1 with default parameters (Browning and Browning, 2016). Ultimately, a total of 2,037,712 SNPs was used for further analysis.

Genetic Parameter Estimations

Before performing genomic prediction, in order to assess how much phenotypic variability could be explained by the genetic variation of the WGS data, the variance components of the startle response and starvation resistance were estimated in the male and female lines, and the variance components were estimated by the information restricted maximum likelihood (REML) method implemented in the LDAK software (Speed and Balding, 2019). The statistical model was

$$\mathbf{y} = \boldsymbol{X}\mathbf{b} + \boldsymbol{Z}\mathbf{g} + \mathbf{e}$$

where y is a vector of the phenotypic values of all individuals; $\mathbf{b}$ is the fixed effect, including the Wolbachia infection status; $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the incidence matrices relating the fixed and polygene effects to the phenotypic records; $\mathbf{g}$ is a vector of the polygene effect of all individuals, which is assumed to be distributed as $\mathbf{g} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\sigma}_{\mathbf{g}}^{\mathbf{2}}\mathbf{G})$; and $\mathbf{e}$ is

the residual term, which is assumed to follow a normal distribution of $\mathbf{e} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\sigma}_{\mathbf{e}}^{\mathbf{2}}\mathbf{I})$. In addition, $\mathbf{G}$ is the standardized relatedness matrix calculated by GEMMA using all SNPs according to (VanRaden, 2008):

$$\mathbf{G} = \frac{\boldsymbol{MM}^{T}}{2\sum_{i=1}^{m} p_i(1-p_i)},$$

where $\mathbf{M}$ is a matrix of centered genotypes, and p_i is the minor allele frequency of $\mathbf{SNP_i}$.

Strategies for selecting the feature markers in genomic prediction

In order to improve the predictive ability of whole genome prediction, four strategies were used to preselect feature SNPs from the WGS data, including the following: 1) SNPs were preselected from the WGS data based on the GWAS results (abbreviation as "S_ GWAS"); 2) SNPs were preselected from the WGS data based on the genome positions of significant genes from the transcriptome-wide association study (abbreviation as "S_ TWAS"); 3) SNPs were preselected from the WGS data based on the results of the eQTL mapping of all genes (abbreviation as "S_eQTL_A"); and 4) SNPs were preselected from the WGS data based on the results of the eQTL mapping of significant genes from TWAS (abbreviation as "S_eQTL_S"). In all scenarios, if no gene or SNP remained after the cut-off thresholds of different categories, the top two genes or five SNPs were exacted as feature markers.

SNPs preselected from the WGS data based on the GWAS results (S_GWAS)

In order to link genomic variation with complex traits, GWASs were performed for each sex separately for the analyzed traits in the training population. Univariate tests of

association were performed using a mixed model approach implemented in the GEMMA v0.98.1 software (Zhou and Stephens, 2014). The model was

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zg} + \mathbf{Sa} + \mathbf{e},$$

where y is a vector of the phenotypic values of individuals in the training set; **a** is the additive effect of the candidate variants to be tested for association; $\mathbf{S}$ is a vector of an SNP; and the other parameters are defined as above. A Wald test was applied to test the alternative hypotheses of each SNP in the univariate models. After the GWAS analysis, the SNPs associated with related traits were divided into different categories based on p values of less than 0.05, 0.001, 0.0001, 0.00001, or 0.000001. Then, the different categories of significant SNPs were extracted from the WGS data as genomic features.

SNPs preselected from the WGS data based on the genome position of significant genes from TWAS (S_TWAS)

In order to link the gene expression level with complex traits, TWASs were performed for each sex separately for the analyzed traits in the training population. The univariate tests of association were performed using a mixed model approach implemented in 'rMVP', a package in R (https://github.com/xiaolei-lab/rMVP). The model was

$$\mathbf{y} = \mathbf{Xb} + \boldsymbol{Z}\mathbf{g} + \boldsymbol{T}\mathbf{a} + \mathbf{e},$$

where y is a vector of the phenotypic values of individuals in the training set; $\boldsymbol{T}$ is a vector of the gene expression level; and the other parameters are defined as above. A Wald test was applied to test the alternative hypotheses of each gene in the univariate models. After the TWAS analysis, the significant gene expression levels associated with

related traits were divided into different categories based on p values of less than 0.05, 0.001, 0.0001, 0.00001, or 0.000001. Then, the SNPs located in significant genes were extracted based on their corresponding genomic positions and combined together as feature markers.

SNPs preselected from the WGS data based on the results of the eQTL mapping of all genes (S_eQTL_A)

In order to link genome variation with the gene expression level, eQTL mapping was performed for each sex separately for each gene expression level using the WGS data. Univariate tests of association were performed using a mixed model approach implemented in the GEMMA v0.98.1 software (Zhou and Stephens, 2014). The model was

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zg} + \mathbf{Sa} + \mathbf{e},$$

where y is a vector of each gene expression level of all individuals; b is the fixed effect, including Wolbachia infection status and five major polymorphic inversions (In2L(t), In2R(NS), In3R(P), In3R(K), and In3R(Mo)); **S** is a vector of a SNP; and the other parameters are defined as above. A Wald test was applied to test the alternative hypotheses of each SNP in the univariate models. After eQTL mapping, the significant eQTLs of each gene were divided into different categories based on p values of less than 0.05, 0.001, 0.0001, 0.00001, or 0.000001. Then, the different categories of significant eQTLs of each gene were extracted from the WGS data and combined together as feature markers.

SNPs preselected from the WGS data based on the results of the eQTL mapping of significant genes (S_eQTL_S)

After the TWAS analysis, the significant gene expression levels associated with related traits were divided into different categories based on a p value less than 0.05, 0.001, 0.0001, 0.00001, or 0.000001. After eQTL mapping, the significant eQTLs of each significant gene were divided into different categories based on a p value 0.05, 0.001, 0.0001, 0.00001, or 0.000001. By combining the results of TWAS and eQTL mapping, the different categories of significant eQTLs for different categories of significant genes were extracted from the WGS data and combined together as feature markers.

Genomic prediction model

The breeding values of the genotyped individuals were estimated via genomic best linear unbiased prediction (GBLUP) (VanRaden, 2008) and a genomic feature BLUP model (Sarup et al., 2016). The statistical model for the GBLUP approaches is

$$\mathbf{y} = \boldsymbol{Xb} + \boldsymbol{Zg} + \boldsymbol{e}$$

where y is a vector of the phenotypic values; b is the fixed effect, including Walachia infection status; and the other parameters are defined as above.

The GFBLUP model was an extended BLUP including two random genetic effects:

$$\mathbf{y} = \boldsymbol{Xb} + \boldsymbol{Z_1 f} + \boldsymbol{Z_2 r} + \boldsymbol{e},$$

where $\boldsymbol{y}$, $\boldsymbol{b}$, $\boldsymbol{u}$, **and** $\boldsymbol{e}$ are the same as GBLUP, $\boldsymbol{f}$ is the vector of the genomic values captured by the genetic markers linked to the genomic feature of interest, following a normal distribution of $\boldsymbol{f} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\sigma_f^2 G_f})$; and $\boldsymbol{r}$ is a vector of genomic values captured by the remaining set of genetic markers, following a normal distribution of

$\boldsymbol{r}\sim\mathbf{N}(\mathbf{0},\boldsymbol{\sigma}_{\boldsymbol{r}}^{\mathbf{2}}\boldsymbol{G}_{\boldsymbol{r}})$. $\boldsymbol{Z_1}$ and $\boldsymbol{Z_2}$ are the incidence matrices relating the additive genetic values ($\boldsymbol{g}$ and $\boldsymbol{f}$) to the phenotypic records**.** $\boldsymbol{G_f}$ and $\boldsymbol{G_r}$ were constructed according to (VanRaden, 2008) using the preselected and remaining markers, respectively.

In this study, the variance components were estimated in the training set using the REML algorithm via the LDAK software (Speed and Balding, 2019). Finally, using the dispersion-n matrices and the variance components, predictions of genetic values were obtained by solving the mixed model equations.

Predictive ability evaluation

The Pearson's correlation and regression coefficients between the predicted genetic values and the true phenotypic values were used to assess the accuracy and bias of genomic prediction. Ten replicates of five-fold cross-validation were used to avoid overfitting in this study. Briefly, the genotyped individuals were randomly divided into five subsets. One subset was selected as the validation set, and the remaining four were used as the reference set. Then the cross-validation process was repeated five times to ensure that each subset was validated once. Finally, the average accuracy values and the bias of genomic prediction for the ten replicates of five-fold cross-validation are reported.

Results

Summary statistics and genetic parameter estimations of the analyzed traits

Before performing genomic prediction, the summary statistics and genetic parameter estimations of the traits were analyzed in the male and female lines, and the detailed

results are shown in Table 1. The results showed that the times of the startle response were similar between the male (average 28.68 seconds; range: 14.13–41.25) and female (average 28.25 seconds; range: 13.38–42.10) lines. However, the times of starvation resistance in the female lines (average 60.43 hours; range: 34.45–106.56) were much longer than those in the male lines (average 45.52 hours; range: 21.28–72.00). The standard deviations were 6.37 and 6.45 for the startle response and 12.61 and 9.40 for starvation resistance in the male and female lines, respectively. The coefficients of variation were 22.21% and 22.83% for the startle response and 20.87% and 20.65% for starvation resistance in the male and female lines, respectively. This indicates that substantial phenotypic variation exists among these traits. Furthermore, the values (standard error) of the heritability estimates were 0.771 (0.191) and 0.691 (0.222) for the startle response and 0.999 (0.083) and 0.999 (0.071) for starvation resistance in the male and female lines, respectively, indicating that they are high-heritability traits. Using likelihood ratio tests, the levels of significance of the heritability estimates were 0.003 and 0.011 for the startle response and 0.0002 and 0.00002 for starvation resistance, indicating a significant genetic contribution to phenotypic variability.

SNPs preselected from the WGS data based on the GWAS results (S_GWAS) with different p-value cutoffs for genomic prediction

Using S_GWAS with different p-value cutoffs, the accuracy values of both GBLUP and GFBLUP are shown in Table 2. When GBLUP was performed with the complete WGS data, the prediction accuracy values were 0.208±0.020 and 0.181±0.022 for the startle response and 0.272±0.017 and 0.307±0.015 for starvation resistance in the female and

male lines, respectively (Table 2). Using S_GWAS with the optimal p-value cutoffs ($p < 0.05$), the accuracy values of GBLUP were 0.186±0.021 and 0.158±0.022 for the startle response and 0.207±0.020 and 0.268±0.020 for starvation resistance in the female and male lines, respectively (Table 2). These accuracy values, however, were still lower than those of GBLUP with the complete WGS data. Furthermore, by using S_GWAS to perform the genomic prediction, the accuracy of GBLUP increased as the p-value cutoffs increased (Table 2). In other words, the accuracy of GBLUP increased as the number of SNPs increased (Table S2). Using S_GWAS with the optimal p-value cutoffs, the accuracy of GFBLUP was much lower than that of GBLUP (Table 2). In addition, there was no obvious trend for the accuracy of GFBLUP using different p-value cutoffs to preselect SNPs. Overall, using S_GWAS provided lower accuracy and a larger bias of genomic prediction compared to using the complete WGS data for both GBLUP and GFBULP (Table 2, Table S1).

SNPs preselected from the WGS data based on the TWAS results (T_GWAS) with different p-value cutoffs for genomic prediction

The accuracy of both GBLUP and GFBLUP using S_TWAS with different p-value cutoffs is shown in Table 3. The results showed that the accuracy of GBLUP using S_TWAS with the optimal p-value cutoffs ($p < 0.05$) was 0.189±0.022 and 0.118±0.022 for the startle response and 0.128±0.017 and 0.196±0.015 for starvation resistance in the female and male lines, respectively (Table 3). However, compared to the complete WGS data, the accuracy of GBLUP cannot be improved by using S_TWAS. In addition, by using S_TWAS to perform genomic prediction, the accuracy of GBLUP always

increased as the p-value cutoffs or number of SNPs increased (Table 3 and Table S4). Compared with the GBLUP with S_TWAS, GFBLUP resulted in a higher accuracy and a smaller bias of genomic prediction, except for the startle response using p-value cutoffs less than 0.05 (Table 3 and Table S3). However, this still did not result in a higher accuracy compared to using the complete WGS data. However, by using p-value cutoffs less than 0.0001 to preselect the SNPs, the accuracy of GFBLUP was equal to the accuracy of GBLUP with the complete WGS data (Table 3), but the bias of GFBLUP was smaller than that of GBLUP with the complete WGS data (Table S3).

SNPs preselected from the WGS data based on the results of the eQTL mapping of all genes (S_eQTL_A) with different p-value cutoffs for genomic prediction

The accuracy of both GBLUP and GFBLUP using S_eQTL_A with different p-value cutoffs is shown in Table 4. The results showed that the accuracy of GBLUP using S_eQTL_A with the optimal p-value cutoffs was 0.243±0.020 and 0.220±0.022 for the startle response and 0.274±0.017 and 0.305±0.015 for starvation resistance in the female and male lines, respectively (Table 4). Compared to GBLUP with S_eQTL_A, GFBLUP resulted in lower prediction accuracy, except for the startle response using p-value cutoffs less than 0.001 in the male lines (Table 4). Furthermore, by using S_eQTL_A, the trends of the accuracy and bias of genomic prediction were different for the startle response and starvation resistance. For the startle response, by using S_eQTL_A with the optimal strategy, the best accuracy values were represented by increases of 19.12% and 21.55% for GBLUP and 10.78% and 19.89% for GFBULP in the female and male lines, respectively, compared to GBLUP with the complete WGS

data (Table 4). However, the biases of genomic prediction with the optimal preselection SNPs were larger than those of the complete WGS data (Table S5). For starvation resistance, lower accuracy and similar biases of genomic prediction were found in the female and male lines, respectively (Table 4 and Table S5). In addition, when the number of SNPs was sufficiently large, the increased number of SNPs decreased the accuracy of GBLUP (Table 4 and Table S6).

SNPs preselected from the WGS data based on the results of eQTL mapping of significant genes (S_eQTL_S) with different p-value cutoffs for genomic prediction

The accuracy of genomic prediction for the startle response and starvation resistance using S_eQTL_S with different p-value cutoffs is shown in Figures 1 and 2, respectively. For the startle response, when we used p-value cutoffs less than 0.05 or 0.001 to select the significant genes for preselecting SNPs, there was higher accuracy than when using GBLUP with the complete WGS data, except when performing GFBLUP on the female lines (Figure 1). The best accuracy values were 0.258±0.019 and 0.237±0.019 for GBLUP and 0.265±0.018 and 0.245±0.020 for GFBLUP in the female and male lines, respectively (Figure 1). Compared to the GBLUP using complete WGS data, the accuracy values represented increases of 24.04% and 30.94% for GBLUP and 27.40% and 35.36% for GFBULP in the female and male lines, respectively (Figure 1). Furthermore, using SNPs preselected with the optimal strategy, the bias of GBLUP was 0.916±0.080 and 0.851±0.079, which are similar to the bias of GBLUP with the complete WGS data in the female (1.113±0.140) and male lines (1.223±0.177), but

larger biases of GFBLUP were found in the female (0.415±0.099) and male (0.324±0.096) lines (Table S6). However, when we used p-value cutoffs less than 0.0001 or 0.00001 to select the significant genes for preselecting SNPs, we achieved lower accuracy than when using the complete WGS data for both GBLUP and GFBLUP. For starvation resistance, using S_eQTL_S with different p-value cutoffs always resulted in higher accuracy than GBLUP using complete WGS data for both GBLUP and GFBLUP (Figure 2). The best accuracy values were 0.437±0.015 and 0.427±0.015 for GBLUP and 0.419±0.016 and 0.390±0.014 for GFBLUP (Figure 2). Compared to GBLUP with the complete WGS data, the accuracy values represented increases of 60.66% and 39.09% for GBLUP and 54.04% and 27.04% for GFBULP in the female and male lines, respectively (Figure 2). Furthermore, by using SNPs preselected with the optimal strategy, the biases of genomic prediction were 0.897±0.064 and 1.217±0.061 for GBLUP and 1.122±0.060 and 1.106±0.062 for GFBLUP in the female and male lines, respectively; these values are similar to or smaller than the biases of GBLUP with the complete WGS data (1.137±0.078 and 1.153±0.065 in the female and male lines, respectively) (Table S6). In addition, the number of SNPs preselected from the WGS data based on the results of the eQTL mapping of significant genes from TWAS are shown in Table S8.

Discussion

In the present study, we determined the impact of different SNP preselection strategies on prediction accuracy using WGS and gene expression level data. To the best of our knowledge, this is the first time that gene expression level data were used to preselect feature SNPs to improve the accuracy of genomic prediction. Overall, using the SNPs preselected from WGS data based on gene expression data results in greater accuracy and a smaller bias of genomic prediction for the startle response and starvation resistance in Drosophila. In particular, by using the SNPs preselected from the eQTL mapping of significant genes, the best accuracy values represented increases of 60.66% and 39.09% for the startle response and 27.40% and 35.36% for starvation resistance in the female and male lines, respectively, compared to GBLUP with the complete WGS data. The new knowledge gained from this study will help scholars enrich the use of omics data to improve the power of genomic prediction.

Total genomic heritability and prediction accuracy

Before performing genomic prediction, the heritability estimates of analyzed traits were estimated in the male and female lines. We found that the heritability of the analyzed traits was very high, especially for starvation resistance, which almost explains the whole phenotypic variability in both the female and male lines (Table 1). These results are similar to those of a previous study (Morgante et al., 2019), but higher than the results in (Li et al., 2019). This may be due to the quality control of the SNPs, the number of lines, and the line means for phenotypes in the present study, which are the

same as those used in (Morgante et al., 2019) and different from those in (Gao et al., 2017). The high heritability of the analyzed traits showed that most loci that affect the traits have additive gene actions or contributions from non-additive gene actions at many loci. If additive gene action contributed to high heritability, high heritability would easily achieve a high prediction accuracy (Daetwyler et al., 2008). However, in this study, the high heritability of traits did not result in a high prediction accuracy. Using WGS data, the accuracy values of GBLUP were 0.208±0.020 (0.181±0.022) for the startle response and 0.272±0.017 (0.307±0.015) for starvation resistance in the female (male) lines (Table 2). One possible reason for this result is that the size of the reference population is very small for genomic predictions. The other possible reason is that non-additive gene actions might contribute to the high estimated additive genetic variation components (Maki-Tanila and Hill, 2014). A previous study found that epistasis dominates the genetic architecture of Drosophila's quantitative traits (Huang et al., 2012). Therefore, the high heritability of the analyzed traits was most likely the result of non-additive gene actions. In addition, the accuracy values of GBLUP in the present study were different than those in (Ober et al., 2012; Edwards et al., 2016) and similar to those in (Morgante et al., 2019). This difference may be due to the quality control of SNPs, the number of lines, fixed effects, the cross-validation procedure, or the size of the reference population.

Genomic feature BLUP model for genomic prediction

GFBLUP is an expansion model for traditional GBLUP that separates the total genomic components into two random genetic components using prior biological knowledge

(Sarup et al., 2016). If a genomic feature contains more causal variants, GFBLUP always has greater accuracy by adding different weights for the genomic features in the model according to the estimated variance components (Edwards et al., 2016; Sarup et al., 2016). Similar results were also found in this study (Figure 1 and Figure 2). Furthermore, the accuracy of GBLUP was influenced by the composition's genomic features. If few (or even no) causal variants are contained in the genomic feature, the accuracy of GFBLUP will decrease because of too much weight being given to spurious genomic features (Fang et al., 2017). Similar results were also found in this study (Table 2). If the genomic feature contains a large proportion of causal variants, the GFBLUP further increases its prediction accuracy compared to GBLUP with genomic features only or the complete WGS data (Edwards et al., 2016; Sarup et al., 2016). For example, when p-values of TWAS and eQTL mapping less than 1e-05 and 0.001 were used to preselect the SNPs as a genomic feature, the accuracy values of GBLUP with the genomic feature were 0.418 and 0.353 for starvation resistance in the female and male lines, respectively; these values are lower than the accuracy of GFBLUP (0.419 and 0.381 for the female and male lines) (Figure 2). However, if the genomic feature almost contains all causal variants, GFBLUP results in a lower accuracy compared to GBLUP with genomic features only. For example, when using SNPs preselected based on the best parameter (the p-values of the TWAS and eQTL mapping were less than 1e-05 and 0.05) as the genomic feature, the accuracy values of GBLUP with the genomic feature were 0.437 and 0.414 for starvation resistance in the female and male lines, respectively, which were higher than the accuracy value of GFBLUP (0.355 and 0.369 for the female

and male lines) (Figure 2). Therefore, the strength of GFBLUP is dependent on the preselection strategy for genomic features.

SNP preselection strategies influencing prediction accuracy

Performing genomic predictions with prior biological knowledge can improve the predictive ability for complex traits (de Los Campos et al., 2013; Edwards et al., 2016; MacLeod et al., 2016). In this study, using the association analysis method, four strategies were proposed to preselect SNPs from WGS data for genomic prediction. We found that using S_GWAS did not improve the prediction accuracy values, especially for p-value cutoffs less than 0.001 (Table 2). Similar results were also indicated in previous studies using SNPs preselected from GWAS (Veerkamp et al., 2016; Ye et al., 2019). The main reason for this result is that overfitting decreases the prediction accuracy. Overfitting means that a small proportion of variants captured a large proportion of variant components in the prediction model (Table S9 and Table S10). In addition, a smaller number of SNPs were preselected based on the p-value of GWAS (Table S2), which is similar to the results of a previous study, which showed that the accuracy of GBLUP decreased as the number of SNPs decreased (Ober et al., 2012). Moreover, using S_TWAS with different p-value cutoffs to perform genomic prediction resulted in lower prediction accuracy values compared to GBLUP with the complete WGS data (Table 3). However, compared with S_GWAS, there are no overfitting problems in the prediction model using S_TWAS (Table S11 and Table S12). The main factor for this decrease in prediction accuracy is that very few causal variants were detected using the genome position of the significant genes from TWAS (Table S4)

because the gene expression level is not only effected by the variants near the regions of this gene (cis-eQTL) but also by the other SNPs in the genome (trans-eQTL) (Everett et al., 2020). This phenomenon was confirmed by the greater accuracy values obtained using the SNPs preselected from the eQTL mapping of significant genes (Figure 1 and Figure 2).

Furthermore, when using S_eQTL_A with different p-value cutoffs to perform genomic prediction, only the startle response produced greater accuracy values compared to GBLUP with the complete WGS data. This is most likely because extreme noise was avoided using eQTL mapping to preselect the SNPs for genomic prediction. Because the expression of numerous genes was found in Drosophila (Everett et al., 2020), combining the significant eQTLs of each gene together almost covered the whole genome (Table S6).

Finally, we combined the strength of TWAS and eQTL mapping by using S_eQTL_S to perform genomic prediction and obtained a higher accuracy and smaller bias of genomic prediction (Figure 1 and Figure 2), as the link between genomic variation and organismal phenotypes could only be determined by TWAS and eQTL mapping using gene expression data (Azodi et al., 2019). Briefly, the significant genes from TWAS in the training population represented the main gene expression level directly associated with the traits, and eQTL mapping of the whole population determined the significant SNPs associated with the gene expression level. In addition, combining the analyses of genomic variation with those of transcriptional variation and organismal phenotype variation allowed us to determine the gene networks associated with complex traits

(Everett et al., 2020) so that the gene–gene interactions (epistasis) associated with complex traits could be captured. Overall, using genomic features preselected from multi-omics data is a feasible strategy to improve the power of genomic prediction.

Challenges for integrating transcriptomic data into genomic predictions

Both this study and several previous studies have indicated that integrating transcriptomic data into genomic prediction is a feasible method to improve the power of genomic prediction (Azodi et al., 2019; Hu et al., 2019; Morgante et al., 2019). However, using transcriptomic data for genomic prediction in animal and plant breeding remains challenging because RNA sequencing thousands of individuals of interest is still too costly for routine implementation, especially in practical breeding. Furthermore, unlike SNP, the level of gene expression is tissue-specific and time-dependent. Hence, the RNA must be extracted from the tissue associated with the trait of interest during the correct periods. However, this is very difficult to achieve in practice. In this study, RNA was extracted from whole flies, which ignored the tissue-specific and time-dependent effect such that the gene expression levels represented the average across all tissues (Everett et al., 2020). It is important to balance the costs and benefits of using transcriptomic information when integrating transcriptomic data into genomic predictions for practical implementations.

Conclusion

Overall, multi-omics data can assist genomic feature preselection and improve the performance of genomic prediction. The new knowledge gained from this study will

enrich the use of multi-omics in genomic prediction.

List of abbreviations

DGRP: Drosophila Genetic Reference Panel; eQTL: expression quantitative trait locus; G: standardized relatedness matrix; GBLUP: genomic best linear unbiased prediction; GFBLUP: genomic feature best linear unbiased prediction; GS: Genomic selection; GWAS: Genome-wide association study; LD: Linkage disequilibrium; MAF: Minor allele frequency; Omics: multiple genome-level; QTL: Quantitative trait locus; REML: restricted maximum likelihood; SNP: Single nucleotide polymorphism; TWAS: transcriptome-wide association study; WGS: Whole genome sequencing;

Availability of data and materials

The WGS data were downloaded from the Drosophila Genetic Reference Panel (DGRP) (http://dgrp.gnets.ncsu.edu/). The mean quantitative trait values and gene expression levels were taken from a previous study (Everett et al., 2020). The gene expression data can be found in GEO (accession GSE117850).

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Natural Science Foundation of China (31772556), the grants from the earmarked fund for China Agriculture Research System (CARS-35), and the Science and Technology Innovation Strategy projects of Guangdong Province (Grant No. 2018B020203002).

Authors' contributions

SPY, ZZ, and JQL conceived the study and designed the project and helped draft. SPY performed genomic prediction and analyzed the accuracy. All authors read and approved the manuscript.

Acknowledgements

The authors are grateful to Prof. Trudy F. C. Mackay et al., who shared the resources of DGRP lines in public dataset.

References

Abdollahiarpanahi, R., Morota, G., and Peñagaricano, F. (2017). Predicting bull fertility using genomic data and biological information. *Journal of Dairy Science* 100(12)**,** 9656.

Azodi, C.B., Pardo, J., VanBuren, R., de Los Campos, G., and Shiu, S.H. (2019). Transcriptome-based prediction of complex traits in maize. *Plant Cell*. doi: 10.1105/tpc.19.00332.

Daetwyler, H.D., Villanueva, B., and Woolliams, J.A. (2008). Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One* 3(10)**,** e3395. doi: 10.1371/journal.pone.0003395.

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 Genetics* 9(7)**,** e1003608. doi: 10.1371/journal.pgen.1003608.

Dimitrakopoulos, L., Prassas, I., Diamandis, E.P., and Charames, G.S. (2017). Onco-proteogenomics: Multi-omics level data integration for accurate phenotype prediction. *Critical Reviews in Clinical Laboratory Sciences* 54(6)**,** 414-432. doi: 10.1080/10408363.2017.1384446.

Edwards, S.M., Sorensen, I.F., Sarup, P., Mackay, T.F.C., and Sorensen, P. (2016). Genomic Prediction for Quantitative Traits Is Improved by Mapping Variants to Gene Ontology Categories in Drosophila melanogaster. *Genetics* 203(4)**,** 1871-+. doi: 10.1534/genetics.116.187161.

Everett, L.J., Huang, W., Zhou, S., Carbone, M.A., Lyman, R.F., Arya, G.H., et al. (2020). Gene expression networks in the Drosophila Genetic Reference Panel. *Genome Res* 30(3)**,** 485-496. doi: 10.1101/gr.257592.119.

Fang, L., Sahana, G., Ma, P., Su, G., Yu, Y., Zhang, S., et al. (2017). Exploring the genetic architecture and improving genomic prediction accuracy for mastitis and milk production traits in dairy cattle by mapping variants to hepatic transcriptomic regions responsive to intra-mammary infection. *Genet Sel Evol* 49(1)**,** 44. doi: 10.1186/s12711-017-0319-0.

Gao, N., Martini, J.W.R., Zhang, Z., Yuan, X.L., Zhang, H., Simianer, H., et al. (2017). Incorporating Gene Annotation into Genomic Prediction of Complex Phenotypes. *Genetics* 207(2)**,** 489-501.

Garcia-Ruiz, A., Cole, J.B., VanRaden, P.M., Wiggans, G.R., Ruiz-Lopez, F.J., and Van Tassell, C.P. (2016). Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection.

Proceedings of the National Academy of Sciences of the United States of America 113(28)**,** E3995-4004. doi: 10.1073/pnas.1519061113.

Guo, Z., Magwire, M.M., Basten, C.J., Xu, Z., and Wang, D. (2016). Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. *Theor Appl Genet* 129(12)**,** 2413-2427. doi: 10.1007/s00122-016-2780-5.

Heidaritabar, M., Calus, M.P., Megens, H.J., Vereijken, A., Groenen, M.A., and Bastiaansen, J.W. (2016). Accuracy of genomic prediction using imputed whole-genome sequence data in white layers. *J Anim Breed Genet* 133(3)**,** 167-179. doi: 10.1111/jbg.12199.

Hu, X., Xie, W., Wu, C., and Xu, S. (2019). A directed learning strategy integrating multiple omic data improves genomic prediction. *Plant Biotechnol J* 17(10)**,** 2011-2020. doi: 10.1111/pbi.13117.

Huang, W., Massouras, A., Inoue, Y., Peiffer, J., Ramia, M., Tarone, A.M., et al. (2014). Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. *Genome Res* 24(7)**,** 1193-1208. doi: 10.1101/gr.171546.113.

Huang, W., Richards, S., Carbone, M.A., Zhu, D., Anholt, R.R., Ayroles, J.F., et al. (2012). Epistasis dominates the genetic architecture of Drosophila quantitative traits. *Proc Natl Acad Sci U S A* 109(39)**,** 15553-15559. doi: 10.1073/pnas.1213423109.

Iheshiulor, O.O., Woolliams, J.A., Yu, X., Wellmann, R., and Meuwissen, T.H. (2016).

Within- and across-breed genomic prediction using whole-genome sequence and single nucleotide polymorphism panels. *Genet Sel Evol* 48(1), 15. doi: 10.1186/s12711-016-0193-1.

Kemper, K.E., Reich, C.M., Bowman, P.J., Vander Jagt, C.J., Chamberlain, A.J., Mason, B.A., et al. (2015). Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy of across-breed genomic predictions. *Genetics Selection Evolution* 47, 29. doi: 10.1186/s12711-014-0074-4.

Li, Z., Gao, N., Martini, J.W.R., and Simianer, H. (2019). Integrating Gene Expression Data Into Genomic Prediction. *Frontiers in Genetics* 10. doi: 10.3389/fgene.2019.00126.

Mackay, T.F.C., Richards, S., Stone, E.A., Barbadilla, A., Ayroles, J.F., Zhu, D.H., et al. (2012). The Drosophila melanogaster Genetic Reference Panel. *Nature* 482(7384), 173-178.

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.

Maki-Tanila, A., and Hill, W.G. (2014). Influence of gene interaction on complex trait variation with multilocus models. *Genetics* 198(1), 355-367. doi: 10.1534/genetics.114.165282.

Meuwissen, T., and Goddard, M. (2010). Accurate Prediction of Genetic Values for

Complex Traits by Whole-Genome Resequencing. *Genetics* 185(2), 623-631. doi: 10.1534/genetics.110.116590.

Meuwissen, T.H.E., Hayes, B.J., and Goddard, M.E. (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics* 157(4), 1819-1829. doi: 10.1017/S0016672301004931.

Morgante, F., Huang, W., Sørensen, P., Maltecca, C., and Mackay, T.F.C. (2019). Leveraging multiple layers of data to predict Drosophila complex traits. *bioRxiv*, 824896.

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(5), e1002685. doi: 10.1371/journal.pgen.1002685.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81(3), 559-575. doi: 10.1086/519795.

Raymond, B., Bouwman, A.C., Schrooten, C., Houwing-Duistermaat, J., and Veerkamp, R.F. (2018). Utility of whole-genome sequence data for across-breed genomic prediction. *Genetics Selection Evolution* 50, 27. doi: 10.1186/s12711-018-0396-8.

Sarup, P., Jensen, J., Ostersen, T., Henryon, M., and Sorensen, P. (2016). Increased prediction accuracy using a genomic feature model including prior information on quantitative trait locus regions in purebred Danish Duroc pigs. *BMC Genet*

17**,** 11. doi: 10.1186/s12863-015-0322-9.

Song, H., Ye, S., Jiang, Y., Zhang, Z., Zhang, Q., and Ding, X. (2019). Using imputation-based whole-genome sequencing data to improve the accuracy of genomic prediction for combined populations in pigs. *Genetics Selection Evolution* 51**,** 58. doi: 10.1186/s12711-019-0500-8.

Speed, D., and Balding, D.J. (2019). SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nature Genetics* 51(2)**,** 277-284. doi: 10.1038/s41588-018-0279-5.

van Binsbergen, R., Calus, M.P., Bink, M.C., van Eeuwijk, F.A., Schrooten, C., and Veerkamp, R.F. (2015). Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. *Genet Sel Evol* 47**,** 71. doi: 10.1186/s12711-015-0149-x.

VanRaden, P.M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science* 91(11)**,** 4414-4423. doi: 10.3168/jds.2007-0980.

Vazquez, A.I., Veturi, Y., Behring, M., Shrestha, S., Kirst, M., Resende, M.F., Jr., et al. (2016). Increased Proportion of Variance Explained and Prediction Accuracy of Survival of Breast Cancer Patients with Use of Whole-Genome Multiomic Profiles. *Genetics* 203(3)**,** 1425-1438. doi: 10.1534/genetics.115.185181.

Veerkamp, R.F., Bouwman, A.C., Schrooten, C., and Calus, M.P. (2016). Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein-Friesian cattle. *Genet Sel Evol* 48(1)**,** 95. doi: 10.1186/s12711-016-0274-1.

Wang, S., Wei, J., Li, R., Qu, H., Chater, J.M., Ma, R., et al. (2019). Identification of optimal prediction models using multi-omic data for selecting hybrid rice. *Heredity* 123(3)**,** 395-406. doi: 10.1038/s41437-019-0210-6.

Xu, Y., Xu, C., and Xu, S. (2017). Prediction and association mapping of agronomic traits in maize using multiple omic data. *Heredity* 119(3)**,** 174-184. doi: 10.1038/hdy.2017.27.

Ye, S., Gao, N., Zheng, R., Chen, Z., Teng, J., Yuan, X., et al. (2019). Strategies for Obtaining and Pruning Imputed Whole-Genome Sequence Data for Genomic Prediction. *Front Genet* 10(673)**,** 673. doi: 10.3389/fgene.2019.00673.

Ye, S., Song, H., Ding, X., Zhang, Z., and Li, J. (2020). Pre-selecting markers based on fixation index scores improved the power of genomic evaluations in a combined Yorkshire pig population. *animal***,** 1-10. doi: 10.1017/s1751731120000506.

Zhang, C., Kemp, R.A., Stothard, P., Wang, Z., Boddicker, N., Krivushin, K., et al. (2018). Genomic evaluation of feed efficiency component traits in Duroc pigs using 80K, 650K and whole-genome sequence variants. *Genet Sel Evol* 50(1)**,** 14. doi: 10.1186/s12711-018-0387-9.

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(3)**,** e93017. doi: 10.1371/journal.pone.0093017.

Zhou, X., and Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat Methods* 11(4)**,** 407-409.

doi: 10.1038/nmeth.2848.

Table

Table 1. Summary statistics and genetic parameter estimations of the analyzed traits

Traits	Startle Response (Seconds)		Starvation Resistance (Hours)	
Sex	Female	Male	Female	Male
[1]N	199	199	198	198
Min	14.13	13.38	34.45	21.28
Max	41.25	42.10	106.56	72.00
Mean	28.68	28.25	60.43	45.52
S.D[2]	6.37	6.45	12.61	9.40
C.V[3]	22.21%	22.83%	20.87%	20.65%
Heritability(S.E[4])	0.771(0.191)	0.691(0.222)	0.999(0.083)	0.999(0.071)
Pval of LRT[5]	0.003	0.011	0.0002	0.00002

[1]N: Number of individuals; [2]S.D: standard deviation; [3]C.V: coefficient of variation; [4] S.E: standard error; Pval of LRT: P-value obtained from likelihood ratio test.

Table 2. The accuracy values of genomic prediction using the SNPs preselected from WGS data based on the GWAS results (S_GWAS) with different p-value cutoffs

Model	P-value cutoffs[1]	Prediction accuracy(Mean±S.E[2])			
		Startle Response		Starvation Resistance	
		Female	Male	Female	Male
GBLUP[4]	All[3]	0.208±0.020	0.181±0.022	0.272±0.017	0.307±0.015
	< 0.05	0.186±0.021	0.158±0.022	0.207±0.020	0.268±0.020
	< 0.001	0.097±0.025	0.087±0.022	0.135±0.025	0.140±0.019
	< 0.0001	0.065±0.018	0.053±0.020	0.100±0.019	0.032±0.021
	< 0.00001	0.066±0.019	0.060±0.025	0.004±0.021	-0.056±0.019
GFBLUP[5]	< 0.05	0.054±0.026	0.049±0.024	0.115±0.025	0.121±0.024
	< 0.001	0.083±0.026	0.034±0.023	0.036±0.025	0.047±0.019
	< 0.0001	0.041±0.020	0.045±0.021	0.130±0.021	0.084±0.020
	< 0.00001	0.068±0.019	0.061±0.025	0.101±0.024	0.045±0.017

[1]P-value cutoffs: using different p-value cutoffs to preselect SNPs from whole genome sequencing (WGS) data based on the results of genome-wide association study (GWAS); [2] S.E: standard error; [3]All: all SNPs of WGS data; [4]GBLUP: genomic best linear unbiased prediction; [5]GFBLUP: genomic feature best linear unbiased prediction.

Table 3. The accuracy values of genomic prediction using the SNPs preselected from WGS data based on the TWAS results (S_TWAS) with different p-value cutoffs

Model	P-value cutoffs[1]	Prediction accuracy(Mean±S.E[2])			
		Startle Response		Starvation Resistance	
		Female	Male	Female	Male
GBLUP[4]	All[3]	0.208±0.020	0.181±0.022	0.272±0.017	0.307±0.015
	< 0.05	0.189±0.022	0.118±0.022	0.128±0.017	0.196±0.015
	< 0.001	0.029±0.023	-0.008±0.026	-0.007±0.021	0.081±0.022
	< 0.0001	0.005±0.022	-0.042±0.019	0.002±0.020	0.024±0.017
	< 0.00001	-0.001±0.022	-0.042±0.019	0.002±0.020	0.004±0.020
GFBLUP[5]	< 0.05	0.176±0.024	0.106±0.024	0.196±0.020	0.287±0.017
	< 0.001	0.168±0.022	0.140±0.022	0.266±0.018	0.291±0.016
	< 0.0001	0.170±0.019	0.137±0.024	0.272±0.017	0.288±0.018
	< 0.00001	0.169±0.019	0.137±0.024	0.272±0.017	0.296±0.016

[1]P-value cutoffs: using different p-value cutoffs to preselect genes based on the results of transcriptome-wide association study (TWAS), then extracted SNPs from whole genome sequencing (WGS) data according corresponding genomic positions of genes; [2] S.E: standard error; [3]All: all SNPs of WGS data; [4]GBLUP: genomic best linear unbiased prediction; [5]GFBLUP: a genomic feature best linear unbiased prediction.

Table 4. The accuracy values of genomic prediction using the SNPs preselected from WGS data based on the results of the eQTL mapping of all genes (S_eQTL_A) with different p-value cutoffs.

Model	P-value cutoffs[1]	Prediction accuracy(Mean±S.E[2])			
		Startle Response		Starvation Resistance	
		Female	Male	Female	Male
GBLUP[4]	All[3]	0.204±0.021	0.181±0.022	0.272±0.017	0.307±0.015
	< 0.001	0.220±0.020	0.178±0.023	0.268±0.017	0.296±0.015
	< 0.0001	0.243±0.020	0.191±0.023	0.278±0.017	0.305±0.015
	< 0.00001	0.241±0.020	0.215±0.022	0.274±0.017	0.300±0.014
	< 0.000001	0.208±0.020	0.220±0.022	0.238±0.018	0.288±0.016
GFBLUP[5]	< 0.001	0.183±0.023	0.181±0.022	0.265±0.017	0.292±0.016
	< 0.0001	0.216±0.021	0.178±0.025	0.265±0.018	0.294±0.015
	< 0.00001	0.226±0.021	0.177±0.024	0.252±0.017	0.276±0.017
	< 0.000001	0.187±0.024	0.217±0.022	0.237±0.02	0.272±0.016

[1]P-value cutoffs: using different p-value cutoffs to preselect SNPs from whole genome sequencing (WGS) data based on the results of expression quantitative trait loci (eQTL) mapping of all genes; [2] S.E: standard error; [3]All: all SNPs of WGS data; [4]GBLUP: genomic best linear unbiased prediction; [5]GFBLUP: a genomic feature best linear unbiased prediction.

Figures

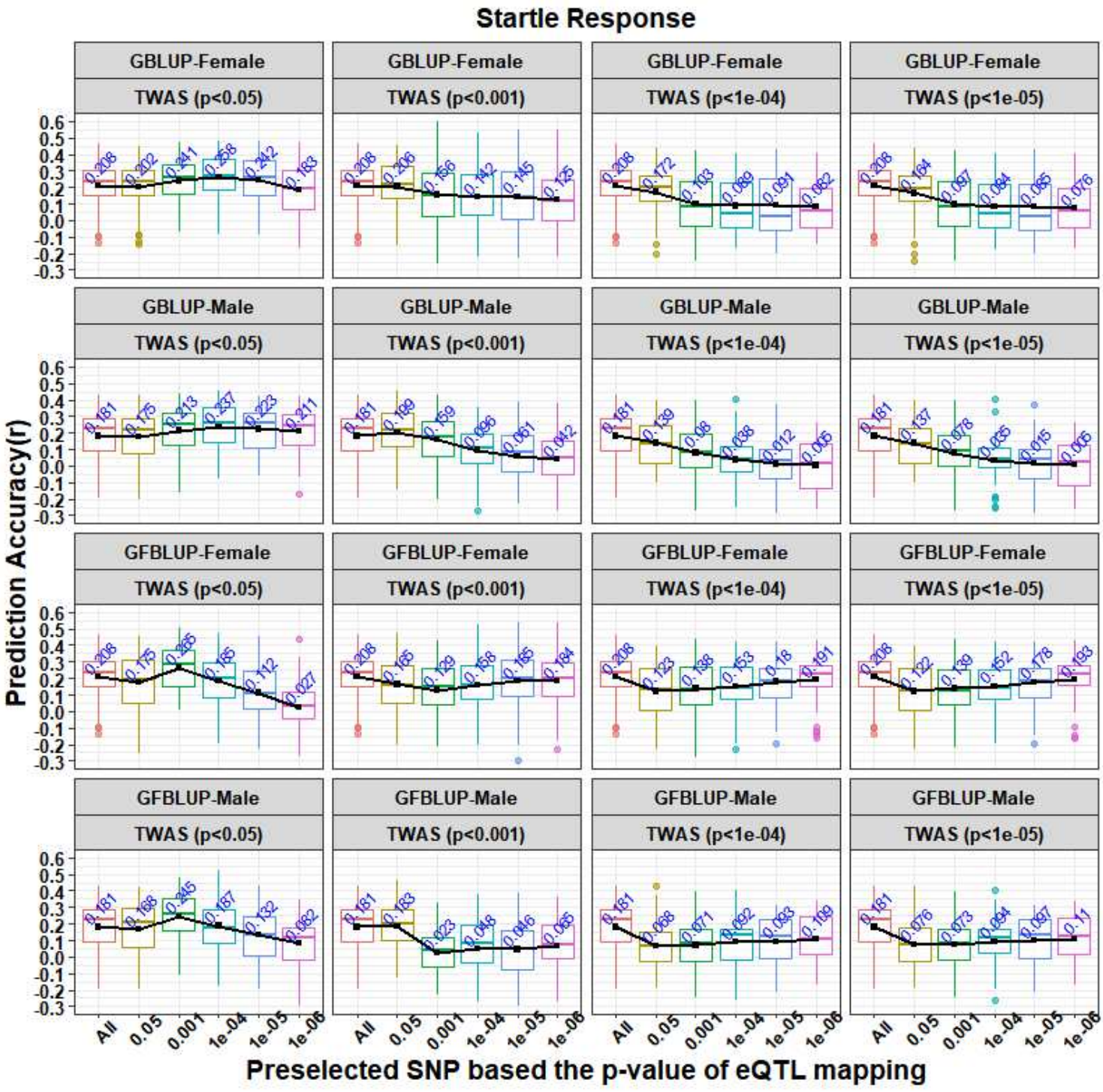

Figure 1

The accuracy values of genomic prediction for the startle response using SNPs preselected from WGS data based on the results of the eQTL mapping of significant genes (S_eQTL_S) with different p-value cutoffs. The Y axis represents the Pearson correlation between the predicted genetic values and the phenotypic values for each trait in the validation sets. Both the X axis and the different colors of box plots

represent the SNP datasets preselected from whole genome sequencing data using different p-value cutoffs based on the results of the eQTL mapping of significant genes from a transcriptome-wide association study (TWAS). GBLUP-Female and GBLUP-Male refer to performing genomic best linear unbiased prediction (GBLUP) on the female and male lines. GFBLUP-Female and GFBLUP-Male refer to performing genomic feature best linear unbiased prediction (GFBLUP) on the female and male lines. TWAS (p<cutoffs) refers to using the p-value cutoffs to preselect significant genes from TWAS. Black lines indicate the trend of the average accuracy in different scenarios.

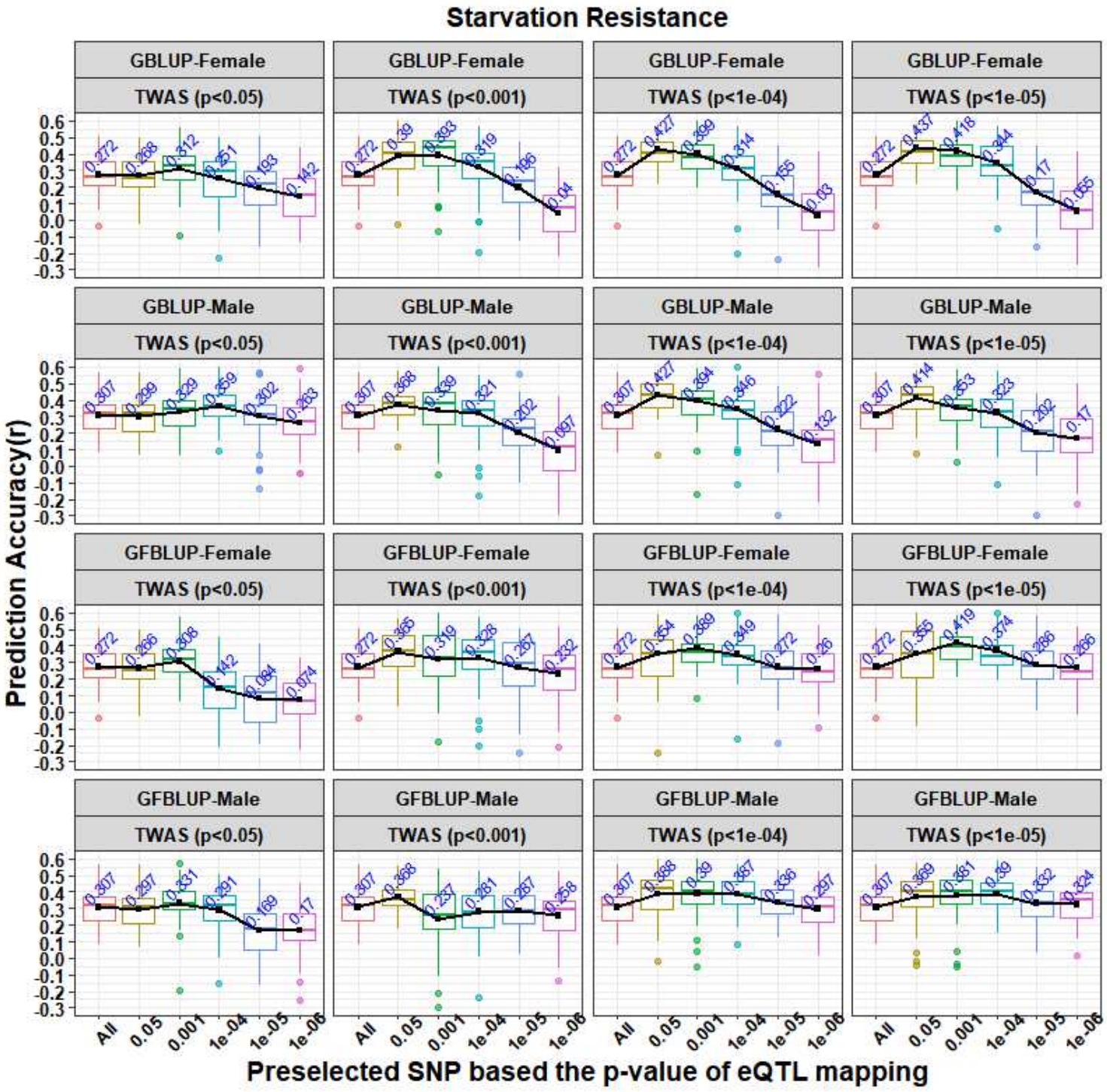

Figure 2

The accuracy values of genomic prediction for the starvation resistance using SNPs preselected from WGS data based on the results of the eQTL mapping of significant genes (S_eQTL_S) with different p-value cutoffs. The Y axis represents the Pearson correlation between the predicted genetic values and the phenotypic values for each trait in the validation sets. Both the X axis and the different colors of box plots represent the SNP datasets preselected from whole genome sequencing data using different p-value cutoffs based on the results of the eQTL mapping of significant genes from a transcriptome-wide association study (TWAS). GBLUP-Female and GBLUP-Male refer to performing genomic best linear unbiased prediction (GBLUP) on the female and male lines. GFBLUP-Female and GFBLUP-Male refer to performing genomic feature best linear unbiased prediction (GFBLUP) on the female and male lines. TWAS (p<cutoffs) refers to using the p-value cutoffs to preselect significant genes from TWAS. Black lines indicate the trend of the average accuracy in different scenarios.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfiles.docx