*Briefings in Bioinformatics*, 22(3), 2021, 1–9

**doi: 10.1093/bib/bbaa103**
Method Review

# Boosting predictabilities of agronomic traits in rice using bivariate genomic selection

Shibo Wang†, Yang Xu†, Han Qu†, Yanru Cui, Ruidong Li, John M. Chater, Lei Yu, Rui Zhou, Renyuan Ma, Yuhan Huang, Yiru Qiao, Xuehai Hu, Weibo Xie and Zhenyu Jia

Corresponding author: Zhenyu Jia, University of California, Riverside, USA. E-mail: shibow@ucr.edu, arthur.jia@ucr.edu
†These authors contributed equally to this work.

## Abstract

The multivariate genomic selection (GS) models have not been adequately studied and their potential remains unclear. In this study, we developed a highly efficient bivariate (2D) GS method and demonstrated its significant advantages over the univariate (1D) rival methods using a rice dataset, where four traditional traits (i.e. yield, 1000-grain weight, grain number and tiller number) as well as 1000 metabolomic traits were analyzed. The novelty of the method is the incorporation of the HAT methodology in the 2D BLUP GS model such that the computational efficiency has been dramatically increased by avoiding the conventional cross-validation. The results indicated that (1) the 2D BLUP-HAT GS analysis generally produces higher predictabilities for two traits than those achieved by the analysis of individual traits using 1D GS model, and (2) selected metabolites may be utilized as ancillary traits in the new 2D BLUP-HAT GS method to further boost the predictability of traditional traits, especially for agronomically important traits with low 1D predictabilities.

**Key words:** bivariate; BLUP; genomic selection; HAT; predictability; metabolites

## Introduction

The advent of cutting-edge technologies has made it feasible to use genome-wide DNA markers to assist in gene discovery, genetic dissection and trait prediction [1], facilitating the rapid selection of superior genotypes and accelerating the breeding cycle. Genomic selection (GS), which was first proposed by Meuwissen *et al.* [2], uses all of the DNA variants across the entire genome to predict complex traits of interest. Advanced methods have been developed for GS analyses, including BLUP [3], LASSO [4, 5], BAYES [2, 6–8], etc., and numerous studies have shown that BLUP-based methods generally outcompete other methods when both prediction accuracy (predictability) and

**Shibo Wang** and **John M. Chater** are postdoc in Dr. Jia's lab.
**Yang Xu** is an assistant professor at Yangzhou University, China.
**Han Qu**, **Ruidong Li**, **Lei Yu** are PhD students at University of California, Riverside, USA.
**Yanru Cui** is an associate professor at Hebei Agricultural University, China.
**Rui Zhou** is a master student in South China University of Technology, China.
**Renyuan Ma** is bachelor student at Bowdoin College, USA.
**Yuhan Huang** is a bachelor student at University of California, Los Angeles, USA.
**Yiru Qiao** is a bachelor student at University of California, Riverside, USA.
**Xuehai Hu** is an associate professor at Huazhong Agricultural University, China.
**Weibo Xie** is a full professor at Huazhong Agricultural University, China.
**Zhenyu Jia** is an associate professor at University of California, Riverside, USA.
**Submitted:** 1 March 2020; **Received (in revised form):** 27 April 2020

computational efficiency are considered [9, 10]. BLUP-HAT [11], a variant of conventional BLUP methodology, further reduces computational burden by avoiding the lengthy cross-validation (CV) procedure, which is routinely used to evaluate the predictive abilities for prediction methods. Due to the complexity of modeling, only univariate (singe-trait or one-dimensional (1D)) GS models have been heavily investigated for these methods [12–20]. However, these widely used 1D GS methods often have low levels of predictability for traits of low heritability. A few studies indicated that multi-trait GS models may increase genomic prediction accuracy compared to the opponent univariate GS models [21–27], but such potential and the trade-off between improvement of predictability and increase in computational burden need further investigation with sufficient data to justify a wider application of multi-trait models in GS.

In this study, we proposed a novel bivariate (2-trait or two-dimensional (2D)) BLUP-HAT GS method to increase trait predictability or accuracy of trait prediction for breeding programs. Another desirable feature of the method is that the HAT method [11] has been incorporated in a 2D BLUP model such that the computational efficiency may be substantially increased. We demonstrated that the new 2D BLUP-HAT GS method outperformed 1D rival GS methods using a large rice dataset, which consists of four traditional traits and 1000 metabolomic traits. The results indicated that (1) the 2D GS analysis generally produces higher predictability than the 1D GS analysis, and (2) traits with low 1D-predictability may significantly benefit from the 2D GS analysis when paired with a carefully selected ancillary trait, for example, a metabolomic trait. The 2D BLUP-HAT GS model may be extended to higher dimensional multivariate models; however, the gain in trait predictability is trivial whereas the increase in computational cost is substantial. We concluded that if data allow, which is often the case, 2D GS analysis should be considered to increase predictability of traits.

## Methods

Suppose $P_s$ is a $n \times 1$ vector for the phenotypic values for trait $s$, where $s = 1$ or 2, and $n$ is the number of individuals in the sample. We use Equation (1) to describe two phenotypes

$$P = \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} = \begin{bmatrix} X & 0 \\ 0 & X \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}, \quad (1)$$

where X is a $n \times q$ design matrix for the fixed effects, $\beta_s$ is a $q \times 1$ vector for the fixed effect for trait $s$, $\xi_s$ is a $n \times 1$ vector for the polygenic effect for trait $s$ with a normal distribution $N(0, K\sigma^2_{A_s})$ and K is the kinship matrix calculated using genomic data [11], and $\varepsilon_s$ is a $n \times 1$ vector representing the residual errors for trait $s$ with a normal distribution $N(0, \sigma^2_s)$. Let $U = \begin{bmatrix} X & 0 \\ 0 & X \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$ and $\xi = \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix}$, then the expectation of the model is

$$E(P) = \begin{bmatrix} X & 0 \\ 0 & X \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = U\beta. \quad (2)$$

Let $G = \begin{bmatrix} \sigma^2_{A_1} & \sigma_{A_{12}} \\ \sigma_{A_{12}} & \sigma^2_{A_2} \end{bmatrix}$ and $R = \begin{bmatrix} \sigma^2_1 & \sigma_{12} \\ \sigma_{12} & \sigma^2_2 \end{bmatrix}$, then the bivariate phenotypic variance is

$$\mathrm{Var}(P) = V = G \otimes K + R \otimes I_n, \quad (3)$$

where $I_n$ is identity matrix of size $n$ and $\otimes$ denotes Kronecker product. The variance components, $\theta = \{\sigma^2_{A_1}, \sigma_{A_{12}}, \sigma^2_{A_2}, \sigma^2_1, \sigma_{12}, \sigma^2_2\}$, can be estimated using the restricted maximum likelihood (REML) method of which the log likelihood function is defined as

$$L(\theta) = -\frac{1}{2}\ln|V| - \frac{1}{2}\ln|U^T V^{-1} U| - \frac{1}{2}\left(P - U\hat{\beta}\right)^T V^{-1}\left(P - U\hat{\beta}\right), \quad (4)$$

where $\hat{\beta} = (U^T V^{-1} U)^{-1}(U^T V^{-1} P)$. The Hendersons mixed model equation becomes

$$\begin{bmatrix} U^T(R \otimes I_n)^{-1} U & U^T(R \otimes I_n)^{-1} I_{2n} \\ I_{2n}(R \otimes I_n)^{-1} U & I_{2n}(R \otimes I_n)^{-1} I_{2n} + (G \otimes K)^{-1} \end{bmatrix} \begin{bmatrix} \beta \\ \xi \end{bmatrix} = \begin{bmatrix} U^T(R \otimes I_n)^{-1} P \\ I_{2n}(R \otimes I_n)^{-1} P \end{bmatrix}. \quad (5)$$

The BLUE and BLUP of the fixed effects and polygenic effect are obtained via

$$\begin{bmatrix} \hat{\beta} \\ \hat{\xi} \end{bmatrix} = \begin{bmatrix} U^T(R^{-1} \otimes I_n) U & U^T(R^{-1} \otimes I_n) \\ (R^{-1} \otimes I_n) U & R^{-1} \otimes I_n + G^{-1} \otimes K^{-1} \end{bmatrix}^{-1} \times \begin{bmatrix} U^T(R^{-1} \otimes I_n) P \\ (R^{-1} \otimes I_n) P \end{bmatrix}. \quad (6)$$

The variance–covariance matrix of the BLUE and BLUP is

$$\mathrm{Var}\begin{bmatrix} \hat{\beta} \\ \hat{\xi} \end{bmatrix} = \begin{bmatrix} U^T(R^{-1} \otimes I_n) U & U^T(R^{-1} \otimes I_n) \\ (R^{-1} \otimes I_n) U & R^{-1} \otimes I_n + G^{-1} \otimes K^{-1} \end{bmatrix}^{-1}. \quad (7)$$

### Prediction of genetic value

Let $P_A$ be the phenotypic values of two traits for $a$ individuals that have been used for developing the bivariate GS model and let $P_B$ be the phenotypic values of two traits for $b$ individuals for which the prediction will be made. We rewrite the model (1) as

$$\begin{bmatrix} P_A \\ P_B \end{bmatrix} = \begin{bmatrix} U_A\beta \\ U_B\beta \end{bmatrix} + \begin{bmatrix} \xi_A \\ \xi_B \end{bmatrix} + \begin{bmatrix} \varepsilon_A \\ \varepsilon_B \end{bmatrix}. \quad (8)$$

The variance–covariance matrix is also partitioned similarly as

$$\mathrm{Var}\begin{bmatrix} P_A \\ P_B \end{bmatrix} = \begin{bmatrix} V_{AA} & V_{AB} \\ V_{BA} & V_{BB} \end{bmatrix} = \begin{bmatrix} G_{AA} & G_{AB} \\ G_{BA} & G_{BB} \end{bmatrix} + \begin{bmatrix} R_{AA} & 0 \\ 0 & R_{BB} \end{bmatrix}, \quad (9)$$

where $G_{AA} = G \otimes K_{AA}$ and $R_{AA} = R \otimes I_a$, then $V_{AA} = G_{AA} + R_{AA}$. Other submatrices are similarly defined. To predict the trait values or genetic values in the test sample, we use the conditional expectation of $p_B$ given $p_A$ (also called BLUP), which is expressed as

$$\begin{aligned} \hat{P}_B &= E(P_B|P_A) \\ &= U_B\hat{\beta} + G_{BA}V_{AA}^{-1}(P_A - U_A\hat{\beta}) \\ &= U_B\hat{\beta} + (G \otimes K_{BA})(G \otimes K_{AA} + R \otimes I_a)^{-1}(P_A - U_A\hat{\beta}) \end{aligned} \quad (10)$$

Let $\xi_B = p_B - U_B\hat{\beta}$ and $\hat{\xi}_B = \hat{p}_B - U_B\hat{\beta}$ be the observed polygenic effect and the predicted polygenic effect, respectively, with

fixed effects being removed. We define $\xi_B = \begin{bmatrix} \xi_{B_1} \\ \xi_{B_2} \end{bmatrix}$, with $\xi_{B_1}$ being the observed polygenic effect for trait 1 and $\xi_{B_2}$ for trait 2, respectively. We also define $\hat{\xi}_B = \begin{bmatrix} \hat{\xi}_{B_1} \\ \hat{\xi}_{B_2} \end{bmatrix}$ as the predicted polygenic effect two traits. The predictabilities for traits 1 and 2 are defined as the squared correlations between the observed polygenic effect and the predicted polygenic effect

$$r^2_{\xi_{B_1}\hat{\xi}_{B_1}} = \frac{\mathrm{Cov}^2\left(\xi_{B_1}, \hat{\xi}_{B_1}\right)}{\mathrm{Var}\left(\xi_{B_1}\right)\mathrm{Var}\left(\hat{\xi}_{B_1}\right)} \text{ and } r^2_{\xi_{B_2}\hat{\xi}_{B_2}} = \frac{\mathrm{Cov}^2\left(\xi_{B_2}, \hat{\xi}_{B_2}\right)}{\mathrm{Var}\left(\xi_{B_2}\right)\mathrm{Var}\left(\hat{\xi}_{B_2}\right)}, \quad (11)$$

respectively.

## 2D BLUP-HAT method

Following the original 1D BLUP-HAT method described by Xu [11], we developed the 2D BLUP-HAT as indicated below. Let $H_R = \hat{\Phi}_A \otimes KV^{-1}$ be the hat matrix [11], then the predicted polygenic effect can be expressed using a linear function of the observed polygenic effect involving the hat matrix, i.e.

$$\hat{\xi} = \begin{bmatrix} \hat{\xi}_1 \\ \hat{\xi}_2 \end{bmatrix} = \begin{bmatrix} \sigma^2_{A_1} & \sigma_{A_{12}} \\ \sigma_{A_{12}} & \sigma^2_{A_2} \end{bmatrix} \otimes KV^{-1}\xi = \hat{\Phi}_A \otimes KV^{-1}\xi = H_R\xi. \quad (12)$$

Let $y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = P$ be the observed predicted phenotypic values for two traits; thus, $\hat{y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \end{bmatrix} = \hat{P} = U\hat{\beta} + \hat{\xi}$ becomes their predicted phenotypic values. Let $\hat{e} = y - \hat{y}$ be the residuals, with $\hat{e}_i = \begin{bmatrix} \hat{e}_{(i)} \\ \hat{e}_{(n+i)} \end{bmatrix}$ as the residuals of two traits for individual i, where $\hat{e}_{(i)}$ is the ith element of the residual vector. The predicted residual for individual i becomes

$$\tilde{e}_i = \begin{bmatrix} \tilde{e}_{(i)} \\ \tilde{e}_{(n+i)} \end{bmatrix} = \left(I_2 - \begin{bmatrix} h_{(i)(i)} & h_{(i)(n+i)} \\ h_{(n+i)(i)} & h_{(n+i)(n+i)} \end{bmatrix}\right)\hat{e}_i, \quad (13)$$

where $h_{(r)(s)}$ represents the single entry on the rth row and the sth column of the hat matrix $H_R$. The total sum of squares of two traits in the bivariate genomicselection is defined as:

$$SS_1 = \sum_{i=1}^{n}\left(y_{1i} - \bar{y}_1\right)^2 \text{ and } SS_2 = \sum_{i=1}^{n}\left(y_{2i} - \bar{y}_2\right)^2 \quad (14)$$

respectively, where $\bar{y}_1 = \frac{\sum_{i=1}^{n} y_{1i}}{n}$ and $\bar{y}_2 = \frac{\sum_{i=1}^{n} y_{2i}}{n}$. The predicted sum of squares for two traits in the bivariate genomic selection is

$$\mathrm{PRESS}_1 = \sum_{i=1}^{n}\tilde{e}^2_{(i)} \text{ and } \mathrm{PRESS}_2 = \sum_{i=1}^{n}\tilde{e}^2_{(n+i)} \quad (15)$$

respectively. The predictabilities of the 2D BLUP-HAT version for two traits in the bivariate genomic selection analysis are

$$r^2_1 = 1 - \frac{\mathrm{PRESS}_1}{SS_1} \text{ and } r^2_2 = 1 - \frac{\mathrm{PRESS}_2}{SS_2} \quad (16)$$

respectively. In the study, we use the 2D BLUP-HAT method in place of the 2D BLUP with a 10-fold cross-validation (BLUP-CV) to increase the computational efficiency of the model.

## Optimization for REML estimation

The eigen decomposition for the kinship matrix is $K = Q\Lambda Q^T$, where Q denotes the eigenvectors (an $n \times n$ matrix) and $\Lambda = \mathrm{diag}\left(\delta_1, \ldots, \delta_n\right)$ denotes the eigenvalues (a diagonal matrix). The bivariate phenotypic variance matrix $\mathrm{var}\begin{bmatrix} P_1 \\ P_2 \end{bmatrix} = V$ can be written as

$$\begin{aligned} V &= \begin{bmatrix} K\sigma^2_{A_1} + I_n\sigma^2_1 & K\sigma_{A_{12}} + I_n\sigma_{12} \\ K\sigma_{A_{12}} + I_n\sigma_{12} & K\sigma^2_{A_2} + I_n\sigma^2_2 \end{bmatrix} \\ &= \begin{bmatrix} Q\Lambda Q^T\sigma^2_{A_1} + QQ^T\sigma^2_1 & Q\Lambda Q^T\sigma_{A_{12}} + QQ^T\sigma_{12} \\ Q\Lambda Q^T\sigma_{A_{12}} + QQ^T\sigma_{12} & Q\Lambda Q^T\sigma^2_{A_2} + QQ^T\sigma^2_2 \end{bmatrix} \\ &= \begin{bmatrix} Q & 0 \\ 0 & Q \end{bmatrix}\begin{bmatrix} \Lambda\sigma^2_{A_1} + I_n\sigma^2_1 & \Lambda\sigma_{A_{12}} + I_n\sigma_{12} \\ \Lambda\sigma_{A_{12}} + I_n\sigma_{12} & \Lambda\sigma^2_{A_2} + I_n\sigma^2_2 \end{bmatrix}\begin{bmatrix} Q^T & 0 \\ 0 & Q^T \end{bmatrix}. \end{aligned} \quad (17)$$

We define $\upsilon_P = \begin{bmatrix} Q^T & 0 \\ 0 & Q^T \end{bmatrix}\begin{bmatrix} P_1 \\ P_2 \end{bmatrix}$ and $\upsilon_U = \begin{bmatrix} Q^T & 0 \\ 0 & Q^T \end{bmatrix}\begin{bmatrix} X & 0 \\ 0 & X \end{bmatrix}$, then the REML function (Equation 4) can be written as

$$\begin{aligned} L'(\theta) = &-\frac{1}{2}\ln\left|V'\right| - \frac{1}{2}\ln\left|\upsilon_U^T\left(V'\right)^{-1}\upsilon_U\right| \\ &-\frac{1}{2}(\upsilon_P - \upsilon_U\beta)^T\left(V'\right)^{-1}(\upsilon_P - \upsilon_U\beta) \end{aligned} \quad (18)$$

where $V' = \begin{bmatrix} V'_{11} & V'_{12} \\ V'_{12} & V'_{22} \end{bmatrix} = \begin{bmatrix} \Lambda\sigma^2_{A_1} + I_n\sigma^2_1 & \Lambda\sigma_{A_{12}} + I_n\sigma_{12} \\ \Lambda\sigma_{A_{12}} + I_n\sigma_{12} & \Lambda\sigma^2_{A_2} + I_n\sigma^2_2 \end{bmatrix}$. The six-variance components are estimated by maximizing the new REML function (Equation 18), in which a novel approach is used for a rapid calculation of $|V'|$ and $(V')^{-1}$ to boost the computational efficiency. Let $\left(V'\right)^{-1} = \begin{bmatrix} W'_{11} & W'_{12} \\ W'_{12} & W'_{22} \end{bmatrix}$, such that $\begin{bmatrix} V'_{11} & V'_{12} \\ V'_{12} & V'_{22} \end{bmatrix}\begin{bmatrix} W'_{11} & W'_{12} \\ W'_{12} & W'_{22} \end{bmatrix} = \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix}$ and eq. (19) as shown at bottom of this page. Note that $V'_{11}$, $V'_{12}$ and $V'_{22}$ are all diagonal matrices and their inverse matrices can be derived without difficulty. For the calculation of $|V'|$, we simply multiply the ith row of $V'$ by $-\frac{V'_{12}[i,i]}{V'_{11}[i,i]}$ and add these values to the jth row of $V'$, where $i \le n$ and $j = i + n$. This change does not affect the result of $|V'|$ but makes the calculation much simpler, i.e.

$$\left(V'\right)^{-1} = \begin{bmatrix} \left(V'_{11} - V'_{12}\left(V'_{22}\right)^{-1}V'_{12}\right)^{-1} & -V'_{12}\left(V'_{22}\right)^{-1}\left(V'_{11} - V'_{12}\left(V'_{22}\right)^{-1}V'_{12}\right)^{-1} \\ -V'_{12}\left(V'_{22}\right)^{-1}\left(V'_{11} - V'_{12}\left(V'_{22}\right)^{-1}V'_{12}\right)^{-1} & V'_{11}\left(V'_{22}\right)^{-1}\left(V'_{11} - V'_{12}\left(V'_{22}\right)^{-1}V'_{12}\right)^{-1} \end{bmatrix} \quad (19)$$

$$
\begin{vmatrix} V'_{11} & V'_{12} \\ V'_{12} & V'_{22} \end{vmatrix} = \begin{vmatrix} \begin{matrix} V'_{11}[1,1] & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & V'_{11}[n,n] \end{matrix} & \begin{matrix} V'_{12}[1,1] & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & V_{12}[n,n]' \end{matrix} \\ \begin{matrix} V'_{12}[1,1] & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & V'_{12}[n,n] \end{matrix} & \begin{matrix} V'_{22}[1,1] & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & V'_{22}[n,n] \end{matrix} \end{vmatrix}
$$

$$
= \begin{vmatrix} \begin{matrix} V'_{11}[1,1] & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & V'_{11}[n,n] \end{matrix} & \begin{matrix} V'_{12}[1,1] & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & V'_{12}[n,n] \end{matrix} \\ \begin{matrix} 0 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 0 \end{matrix} & \begin{matrix} V'_{22}[1,1] - V'_{12}[1,1] \times \frac{V'_{12}[1,1]}{V'_{11}[1,1]} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & V'_{22}[n,n] - V'_{12}[n,n] \times \frac{V'_{12}[n,n]}{V'_{11}[n,n]} \end{matrix} \end{vmatrix}
$$

$$
= V'_{11}[1,1] \times \cdots \times V'_{11}[n,n] \times \left( V'_{22}[1,1] - V'_{12}[1,1] \times \frac{V'_{12}[1,1]}{V'_{11}[1,1]} \right) \times \cdots \times \left( V'_{22}[n,n] - V'_{12}[n,n] \times \frac{V'_{12}[n,n]}{V'_{11}[n,n]} \right) \quad (20)
$$

## Rice data

The population used in this study consists of 1619 bins inferred from approximately 270 000 SNPs of the rice genome and 210 $F_9$ recombinant inbred lines (RILs) derived by single-seed descendent from a cross between 'Zhenshan 97' and 'Minghui 63', which are the parents of 'Shanyou 63'—an elite rice hybrid that has been widely cultivated in the last three decades in China. We analyzed four traits: yield (YIELD), 1000-grain weight (KGW), grain number per plant (GRAIN) and tiller number per plant (TILLER). Each trait was measured from four replicated experiments (1997 and 1998 from one location, 1998 and 1999 from another location). In each replicated experiment, eight plants were sampled from each line and the average trait value was calculated and used as the phenotypic value for this line in this experiment. We also analyzed 1000 metabolomic traits, including 683 metabolites measured from flag leaves and 317 metabolites measured from germinated seeds. Two biological replicates were sampled for flag leaves in 2009, while each biological replicate was sampled in 2009 and 2010, respectively, for germinated seeds. Metabolomic data for both tissueshad been log2-transformed to satisfy the normality assumption. The average of two replicates for each metabolite was used for analysis.

## Simulated data

We adopted a series of simulated datasets (supporting material File S1) from a previous study where multivariate GS models were investigated [22]. In the default simulation scenario, a total of 20 SNPs were randomly selected as QTL, the effects of which on two phenotypic traits were sampled from a standard bivariate normal distribution with a correlation of 0.5. The true breeding value for an individual was the sum of these QTL effects for each trait. Random noise generated from a normal distribution was added to achieve the heritability of 0.1 for trait 1 and heritability of 0.5 for trait 2. The covariance of errors between two traits was set to be zero. For each of other simulation scenarios, a single simulation parameter was perturbed at a time from the default scenario. The perturbed parameters included heritability for each trait (0.1, 0.5 and 0.8), genetic correlation between traits (0.1, 0.3, 0.5, 0.7 and 0.9) and number of QTL (20 and 200). Each simulation scenario was repeated 24 times to calculate predictabilities for each comparison between 1D GS and 2D GS analyses.

## Results

We first leveraged the 'big data' of 1000 metabolomic traits [28] to conduct GS analysis using the genotypic data of 1619 genetic bins [29]. For each metabolomic trait, the 1D predictability and a total of 999 2D predictabilities (paired with each of the other metabolomic traits) were calculated using the 1D and 2D BLUP-HAT methods, respectively. Then these predictability values were transformed to generate a heat map (Figure 1). Prior to the data transformation, the $j$th row of the heat map represents the 1000 predictabilities calculated for the $j$th metabolomic trait, where $j = 1, \ldots, 1000$. Note that the diagonal entry in the heat map, i.e. element $[j, j]$, represents the 1D predictability calculated for the $j$th metabolomic trait using the 1D BLUP-HAT method; while the element $[j, k]$, an off-diagonal entry in the heat map, denotes the 2D predictability of the $j$th metabolomic trait when it is analyzed with the $k$th metabolomic trait using the 2D BLUP-HAT method. The 1000 metabolomic traits (or the 1000 rows in the heat map) have been sorted based on the 1D predictabilities

by descending order, i.e. metabolomic traits of higher 1D predictabilities are placed on the top of the heat map and *vice versa*. Then, the 1000 values in each row (one 1D predictability +999 2D predictabilities for each metabolomic trait) were transformed to their ranks followed by a division by 1000, yielding standardized rank ratio values ranging from 0 to 1 (yellow to blue color scale in the final heat map) to represent the relative low/high predictabilities for that metabolomic trait. Hierarchical clustering on the columns of the heat map disclosed the 'principle' metabolites (denoted by the vertical blue stripes in Supplementary Figure S1) that generally boosted predictabilities of other metabolomic traits when they were paired with these principle metabolites. When paired with a traditional trait (YIELD, KGW, GRAIN or TILLER), some of these principle metabolites (marked by the red vertical bars above the heat map in Supplementary Figure S1) significantly improved the predictabilities of the traditional trait in the 2D GS analysis as compared to the 2D GS analysis where the subject traditional trait was paired with any other traditional trait. These 'celebrity' metabolites identified in the 2D GS analysis warrant further investigation to disclose their biological roles that substantially contribute to the increase in the predictabilities of traditional traits in rice.

Two color bars (1 and 2) on the left of the heat map in Figure 1 show the advantages of 2D over 1D GS methods. Bar 1 denotes the rank ratio values for the predictabilities calculated using 1D BLUP-HAT (diagonal entries extracted from the heat map) for 1000 sorted metabolomic traits. The majority entries in bar 1 are either green or yellow, suggesting that the maximum predictability (denoted by blue in each row) for any metabolomic trait was achieved by the 2D GS analysis rather than the 1D GS analysis. Moreover, bar 1 shows a top-down decrease (from green to yellow) in the rank ratio values, indicating that the advantage of the 2D GS analysis over the 1D GS analysis becomes more apparent in metabolomic traits with low 1D predictabilities. Compared to the 1D predictabilities for the 1000 metabolites, the average gain in predictabilities (measured in percentage) achieved by the 2D GS method is presented in bar 2 (from grey to dark grey, indicating a range of 0–5000%, respectively). Bar 2 shows a top-down increase with horizontal lines (left of bar 2) marking insignificant differences between 2D and 1D predictabilities based on a one-sample *t*-test with significance level of 0.05. The results of this comparison suggested that (1) 1D predictabilities for most metabolomic traits have been significantly improved by 2D GS analysis, and (2) 2D GS analysis is inclined to benefit the traits with low 1D-predictabilities more than the traits with high 1D predictabilities. Although predictability values for two metabolomic traits (labelled by red within bar 2) appeared to be lower in 2D GS analysis when compared to 1D GS analysis, the differences were not statistically significant. Note that 128 metabolomic traits at the bottom of bar 1 are labelled with white because 1D BLUP-HAT GS method failed to estimate non-negative predictability values for them. These results for the 128 metabolites shown in bar 1 are in agreement with the results of the conventional 1D BLUP (bar 3) and 1D LASSO (bar 4) through 10-fold cross-validation, where purple/white in these two color bars represent positive/negative correlations between the observed metabolomic trait values and the predicted metabolomic trait values. These non-positive 1D predictabilities may be due to the indirect or intricate connections between the genomic data and the trait data for single metabolites, which cannot be picked up by 1D GS analysis but may be remedied by 2D GS analysis. For example, predictabilities for more than 100 metabolomic traits that failed in any 1D methods can be successfully estimated when they are paired with other metabolomic traits in 2D BLUP-HAT GS analysis, implying one major advantage of 2D GS analysis over 1D GS analysis.

We also compared the 2D predictabilities with the 1D predictabilities in the analysis of four traditional traits, i.e. YIELD, KGW, GRAIN and TILLER. Three 2D predictabilities were calculated for each trait when this trait was paired with the other three traits, respectively. These results are shown in Supplementary Table S1. Figure 2 shows that the average of 2D predictabilities (indicated by the blue dashed line) is higher than the 1D predictability (indicated by the black dashed line) for every trait, demonstrating again the advantage of 2D GS analysis over 1D GS analysis. This was also supported by the results of the analysis in which each of these four traditional traits was paired with each of the 1000 metabolomic traits using the 2D GS setting (Figure 2). Similarly, YIELD, which had the lowest 1D predictability among four traditional traits, appeared to benefit most from the 2D GS analysis. The predictabilities of the four traditional traits may be further boosted by certain metabolites (denoted by red dots above the blue dashed lines) than by pairing with other traditional traits in 2D GS setting, suggesting that potential metabolites may be identified as ancillary traits to increase agronomically important traits through 2D GS analysis. In Figure 2, for example, the greatest 2D predictability achieved for YIELD through a booster metabolite was 0.235, which had been increased by 18% when compared to 0.199, the 1D predictability for YIELD. Common metabolites have been found to augment predictabilities for any two traits (red triangles) or any three traits (red squares), which indicates the potential biological connections between these traits and these pivotal metabolites and provides guidance for further investigation into the underlying genetic architectures or biochemical pathways of agronomic crops. On the other hand, the metabolomic traits also benefited from the 2D GS when compared with 1D GS (Supplementary Figure S2), indicating a 'win-win' mutual gain from a multivariate analysis.

Finally, we tested our hypothesis that 2D GS outperforms 1D GS by analyzing a series of *in silico* datasets generated by a previous study [22]. The 2D predictabilities were always greater than the 1D predictabilities for the two traits in all combinations of the predefined simulation parameters (Supplementary Table S2). The results of simulation analysis also demonstrated that the low-heritability traits benefit from 2D GS more than the high-heritability traits, and this was true when genetic correlation increased between traits.

## Discussion

It is common to include a number of related traits in a breeding program selection plan. But these traits are often analyzed separately in the GS analysis using univariate models because multivariate GS models are complex and usually inefficient in computation, which explains why multivariate GS models have been rarely used and understudied. The hypothesized benefits of using multivariate GS analysis over univariate models need to be tested using sufficient data. In this study, we leveraged a large dataset, which consists of 4 traditional traits, 1000 metabolomic traits, as well as simulated datasets for three genetic scenarios to demonstrate the advantage of 2D GS analysis over 1D GS analysis. The results showed that (1) the predictabilities for any two traditional traits obtained from 2D analysis were always higher than those achieved by analyzing these two traits separately with 1D analysis. (2) The same conclusion was made when the same comparisons between 2D and 1D analyses were performed

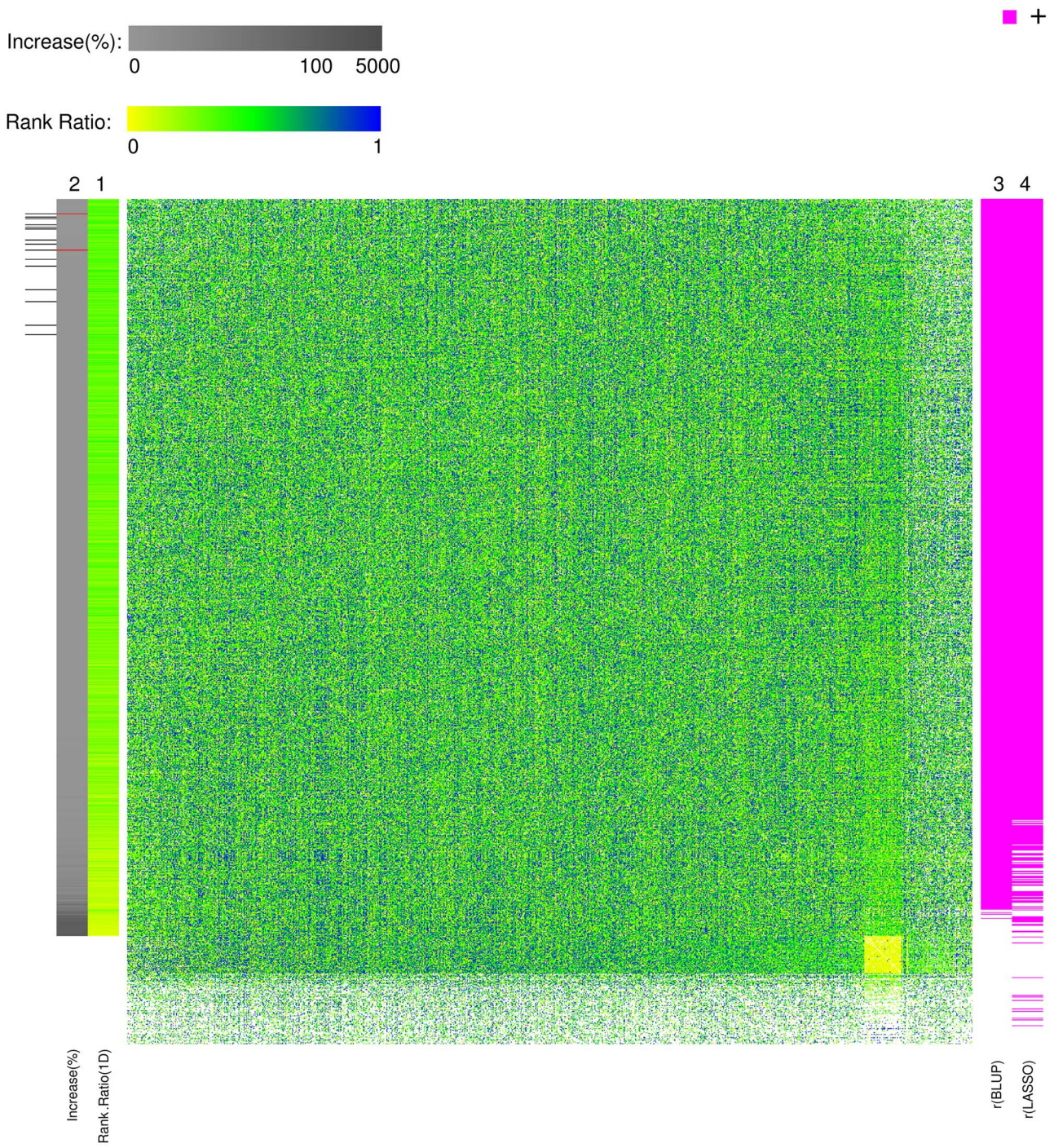


**Figure 1**. Heat map of rank ratios for the 1000 metabolomic traits in 2D GS *versus* 1D GS. Color bar 1 denotes the rank ratio values for the predictabilities calculated using 1D BLUP-HAT (diagonal entries extracted from the heat map) for 1000 sorted metabolomic traits. Color bar 2 represents the increase in predictability when the average of 2D predictabilities was compared with 1D predictability by one-sample *t*-test. The horizontal lines on the left of color bar 2 mark the tests with non-significant results. Color bars 3 and 4 represent the results from conventional 1D BLUP and 1D LASSO, respectively, with purple/white representing positive/negative correlations between the observed trait values and the predicted trait values (see Results for detailed descriptions of figure legends).

using the 1000 metabolomic traits. (3) In the 2D analysis, selected metabolites increased the predictability of any paired traditional trait more than any other traditional trait.

As the end products of regulation at the genomic, transcriptomic and proteomic levels, metabolites serve as the most feasible and direct correlative measure of cellular phenotypes. The rapid development of high-throughput metabolite profiling technologies enable an accurate identification and relative quantification of a large number of metabolites, which has advanced our understanding of the genetic flow from DNA to agronomic traits of interest [30, 31]. A comparison of trait-prediction models using various omics datasets in rice showed that the predictabilities of YIELD and GRAIN based on metabolomic data were generally higher than those based

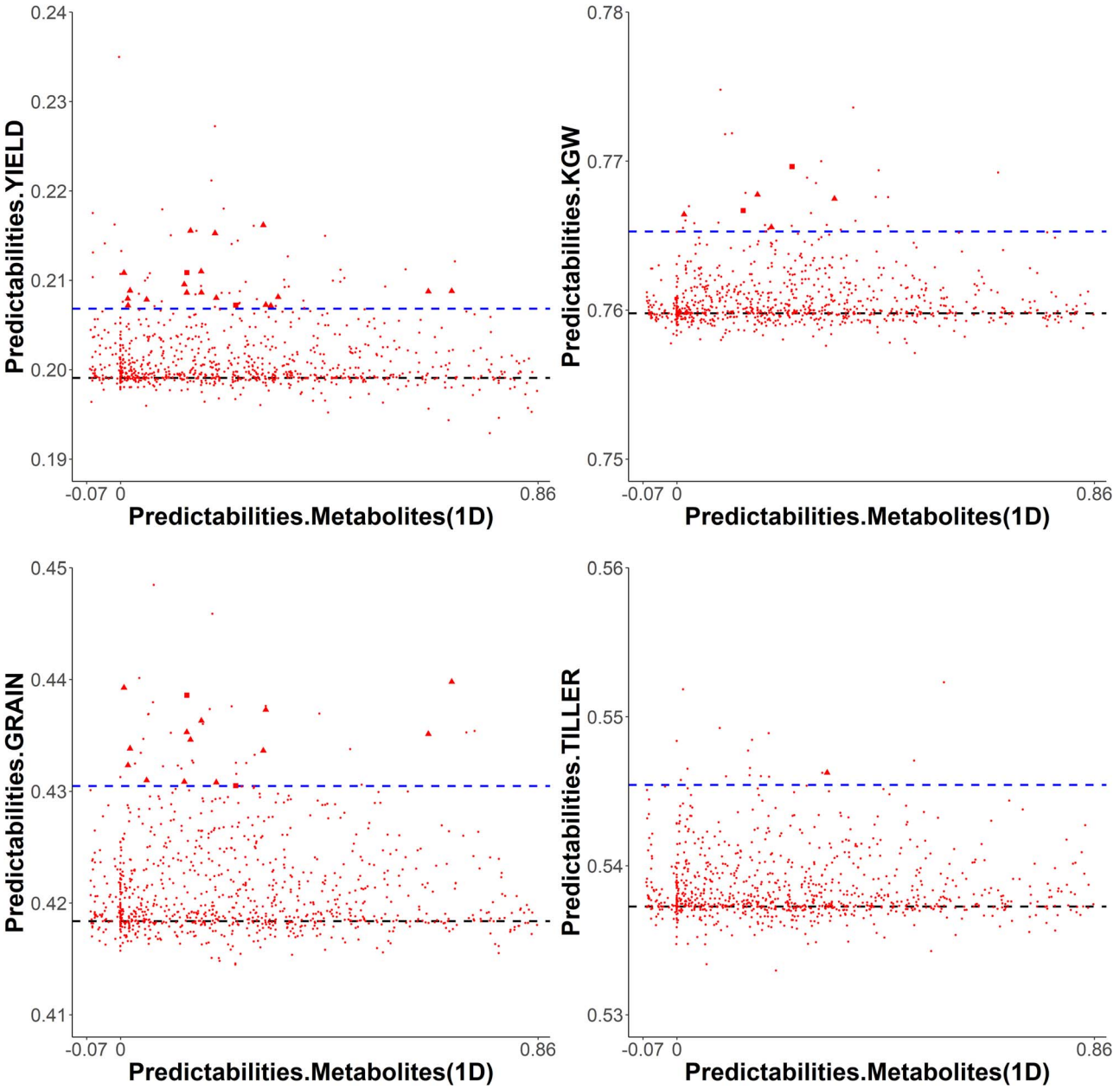


**Figure 2**. The 2D BLUP-HAT GS analysis of each traditional trait paired with each of 1000 metabolomic traits. In each panel, the x-axis indicates the predictability values of the 1000 metabolites estimated in 1D BLUP-HAT GS, and the y-axis indicates the 2D predictability value of the traditional trait when it is analyzed with each of the 1000 metabolites. In each plot, the black dashed line represents the predictability of the traditional trait estimated by 1D BLUP-HAT GS and the blue dashed line represents the average of three 2D predictabilities when this trait was paired with each of 3 other traits at a time. The triangle and square dots in each plot indicate the metabolites that can boost predictabilities for two and three traditional traits, respectively, in 2D BLUP-HAT GS analysis.

on genomic data [9, 10]. This might be due to the fact that, as downstream products, constituents of the metabolome may represent an integral effect of genetic processes at multiple levels and their interactions, which is more correlative to the agronomic traits of interest than the upstream DNA variants are. On the other hand, the predictabilities based on metabolites were lower than those based on genomic data for KGW and TILLER, which suggested that these two traits may be primarily determined by genetic factors at the DNA level. This argument has been supported by the observation that more elite metabolites have been identified for YIELD and GRAIN than for KGW and TILLER. As a result, a multivariate GS analysis including relevant metabolites leverages additional information to boost predictability for the traditional traits of interest, which are also included in the multivariate model.

Our results indicated that traits with relatively low 1D GS predictabilities, such as YIELD in rice, can substantially benefit from 2D GS analysis. For a low GS-predictability trait, as the name would suggest, its variability can be partially explained by genomic data likely due to the weak or indirect connections between this trait and DNA variants. Rather, incorporating another ancillary trait (for example, a selected metabolomic trait), which captures information beyond genomic data, may substantially contribute to the prediction of the target trait in 2D GS setting. The new 2D GS method can help identify 'elite' metabolites, which can increase the predictabilities for

multiple traditional traits such that key biological networks involving genomic loci, metabolites and traits will be discovered. In breeding programs, these elite metabolites may be utilized as candidate ancillary tools for a precision selection, particularly useful to the traits with low 1D GS predictabilities.

We have developed a computationally efficient 2D BLUP-HAT GS methodology to facilitate a wider application of this novel methodology in research and practical breeding. We adopted the HAT method for the 2D BLUP GS method to directly calculate the 'predicted sum of squares' for each trait rather than using the lengthy cross-validation. As a result, the computational efficiency has been substantially increased, which makes bivariate (2D) or higher dimensional GS analysis feasible. A previous study on GS demonstrated that the trait predictabilities calculated using HAT method is very close to those calculated using leave-one-out cross-validation [11], and the closeness between two approaches depends on the size of the training sample. Our proposed 2D GS HAT model can be easily expanded to higher dimensional models to analyze more than two traits; however, the possible gain in trait predictability, which is likely to be miniature, may be compromised by a major increase in computational burden.

Previous studies claimed that the prediction accuracy for a low-heritability trait may be significantly increased in multivariate GS analysis if other correlated high-heritability traits were also used [21, 23, 25, 27, 32]. Our results, which were based on the analysis of a large number of traits, indicated that this is not always the case because metabolites with either higher heritability or higher correlation with a traditional trait, or both, were not consistently effective in boosting the predictability of that trait in the 2D GS analysis (Supplementary Figure S3).

It has been noticed that negative predictability values were incurred for a small number of metabolites, even when the new 2D BLUP-HAT method has been used. This may be explained by the following two reasons. (1) The current model is based on simple linear regression, which may not be able to capture the genetic effects due to the interactions of higher order. (2) The current parameter estimation algorithm implemented in R may not be optimal. To break through these limitations will be our goal for developing the next version of the 2D BLUP-HAT method.

Genome-wide association studies (GWAS) aim to identify genomic loci that are associated with traits of interest. Significant loci identified by GWAS may be used as fixed effects to help increase the trait predictability in GS. For example, He et al. [33] indicated that the combined quantitative trait nucleotides (QTNs) identified from single-locus and multi-locus GWAS approaches (including mrMLM [34]) improved the accuracy in GS analysis. Similar to GS, multivariate GWAS has advantages over univariate GWAS for the same reason that a joint analysis of multiple traits considers covariance between these traits and therefore provides more information for statistical inferences [35–37]. Nevertheless, multivariate analysis includes more parameters and substantially increases the computational burden when compared with univariate analysis. Thus, there is a strong impetus to develop efficient algorithms for multivariate GS as well as for multivariate GWAS.

**Key Points**

- We leveraged 'big' multi-scale rice data to demonstrate that the bivariate genomic selection (GS) analysis generally produces higher predictabilities for two traits than those achieved by the analysis of individual traits using the univariate GS models.
- Selected metabolites may be utilized as ancillary traits in the new bi-variate GS analysis for precision selection, especially for agronomically important traits with low predictabilities in univariate GS analysis, such as yield in rice.
- These elite metabolites identified using the new methodology will help uncover key biological networks involving genomic loci, metabolites and traits, advancing our knowledge of trait-associated genetics.

## Funding

The work was supported by UC Riverside Faculty Start-up Fund, UC Riverside Hellman Fellowship, UC Academic Senate Regents Faculty Fellowship and Faculty Development Award, UC Cancer Research Coordinating Committee Competition Award, and USDA NIFA FACT 2019-67022-29930 to Zhenyu Jia.

## References

1. Araus JL, Kefauver SC, Zaman-Allah M, *et al*. Translating high-throughput phenotyping into genetic gain. *Trends Plant Sci* 2018;**23**:451–66.
2. Hayes B, Goddard M. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 2001;**157**:1819–29.
3. Henderson CR. Best linear unbiased estimation and prediction under a selection model. *Biometrics* 1975;**31**:423–47.
4. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc B Methodol* 1996;**58**:267–88.
5. Asins M. Present and future of quantitative trait locus analysis in plant breeding. *Plant Breedi* 2002;**121**:281–91.
6. De Los CG, Naya H, Gianola D, *et al*. Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* 2009;**182**:375–85.
7. Habier D, Fernando RL, Kizilkaya K, *et al*. Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformat* 2011;**12**:186.
8. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genet* 2013;**9**:e1003264.
9. Wang S, Wei J, Li R, *et al*. Identification of optimal prediction models using multi-omic data for selecting hybrid rice. *Heredity* 2019;**1**.
10. Xu S, Xu Y, Gong L, *et al*. Metabolomic prediction of yield in hybrid rice. *Plant J* 2016;**88**:219–27.
11. Xu S. Predicted residual error sum of squares of mixed models: an application for genomic prediction, G3: genes, genomes. *Genetics* 2017;**7**:895–909.
12. Calus M, Veerkamp R. Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. *J Anim Breed Genet* 2007;**124**:362–8.
13. Legarra A, Robert-Granié C, Manfredi E, *et al*. Performance of genomic selection in mice. *Genetics* 2008;**180**:611–8.
14. Lorenzana RE, Bernardo R. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theor Appl Genet* 2009;**120**:151–61.

15. Daetwyler H, Hickey J, Henshall J, *et al*. Accuracy of estimated genomic breeding values for wool and meat traits in a multi-breed sheep population. *Anim Prod Sci* 2010;**50**:1004–10.
16. Liu Z, Seefried FR, Reinhardt F, *et al*. Impacts of both reference population size and inclusion of a residual polygenic effect on the accuracy of genomic prediction. *Genet Sel Evol* 2011;**43**:19.
17. Gao H, Su G, Janss L, *et al*. Model comparison on genomic predictions using high-density markers for different groups of bulls in the Nordic Holstein population. *J Dairy Sci* 2013;**96**:4678–87.
18. Wolc A, Arango J, Settar P, *et al*. Persistence of accuracy of genomic estimated breeding values over generations in layer chickens. *Genet Sel Evol* 2011;**43**:23.
19. Su G, Christensen OF, Janss L, *et al*. Comparison of genomic predictions using genomic relationship matrices built with different weighting factors to account for locus-specific variances. *J Dairy Sci* 2014;**97**:6547–59.
20. Jia Z. Controlling the Overfitting of heritability in genomic selection through cross validation. *Sci Rep* 2017;**7**:13678.
21. Calus MP, Veerkamp RF. Accuracy of multi-trait genomic selection using different methods. *Genet Sel Evol* 2011;**43**:26.
22. Jia Y, Jannink J-L. Multiple-trait genomic selection methods increase genetic value prediction accuracy. *Genetics* 2012;**192**:1513–22.
23. Guo G, Zhao F, Wang Y, *et al*. Comparison of single-trait and multiple-trait genomic prediction models. *BMC Genet* 2014;**15**:30.
24. Cheng H, Kizilkaya K, Zeng J, *et al*. Genomic prediction from multiple-trait bayesian regression methods using mixture priors. *Genetics* 2018;**209**:89–103.
25. Jiang J, Zhang Q, Ma L, *et al*. Joint prediction of multiple quantitative traits using a Bayesian multivariate antedependence model. *Heredity* 2015;**115**:29.
26. He D, Kuhn D, Parida L. Novel applications of multitask learning and multiple output regression to multiple genetic trait prediction. *Bioinformatics* 2016;**32**:i37–43.
27. Hayashi T, Iwata H. A Bayesian method and its variational approximation for prediction of genomic breeding values in multiple traits. *BMC Bioinformat* 2013;**14**:34.
28. Gong L, Chen W, Gao Y, *et al*. Genetic analysis of the metabolome exemplified using a rice population. *Proc Natl Acad Sci* 2013;**110**:20320–5.
29. Yu H, Xie W, Wang J, *et al*. Gains in QTL detection using an ultra-high density SNP map based on population sequencing relative to traditional RFLP/SSR markers. *PloS One* 2011;**6**:e17595.
30. Matsuda F, Okazaki Y, Oikawa A, *et al*. Dissection of genotype–phenotype associations in rice grains using metabolome quantitative trait loci analysis. *Plant J* 2012;**70**:624–36.
31. Chen W, Wang W, Peng M, *et al*. Comparative and parallel genome-wide association studies for metabolic and agronomic traits in cereals. *Nat Commun* 2016;**7**:1–10.
32. Jia Y, Jannink J-L. Multiple trait genomic selection methods increase genetic value prediction accuracy. *Genet 2012: Genet* **112**:144246.
33. He L, Xiao J, Rashid KY, *et al*. Evaluation of genomic prediction for pasmo resistance in flax. *Int J Mol Sci* 2019;**20**:359.
34. Wang S-B, Feng J-Y, Ren W-L, *et al*. Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci Rep* 2016;**6**:19444.
35. Turley P, Walters RK, Maghzian O, *et al*. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat Genet* 2018;**50**:229–37.
36. Bottolo L, Chadeau-Hyam M, Hastie DI, *et al*. GUESS-ing polygenic associations with multiple phenotypes using a GPU-based evolutionary stochastic search algorithm. *PLoS Genet* 2013;**9**:e1003657.
37. Korte A, Vilhjálmsson BJ, Segura V, *et al*. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat Genet* 2012;**44**:1066.