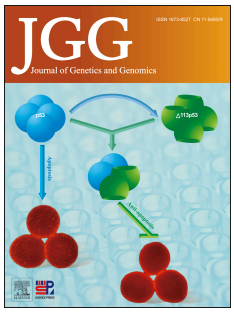
Journal Pre-proof



Multi-view BLUP: a promising solution for post-omics data integrative prediction

Bingjie Wu, Huijuan Xiong, Lin Zhuo, Yingjie Xiao, Jianbing Yan, Wenyu Yang

PII: S1673-8527(24)00332-1

DOI: https://doi.org/10.1016/j.jgg.2024.11.017

Reference: JGG 1431

To appear in: *Journal of Genetics and Genomics*

Received Date: 23 July 2024

Revised Date: 27 November 2024

Accepted Date: 27 November 2024

Please cite this article as: Wu, B., Xiong, H., Zhuo, L., Xiao, Y., Yan, J., Yang, W., Multi-view BLUP: a promising solution for post-omics data integrative prediction, *Journal of Genetics and Genomics*, https://doi.org/10.1016/j.jgg.2024.11.017.

Multi-view BLUP: a promising solution for post-omics data integrative prediction

Bingjie Wu [a,1], Huijuan Xiong [b,1], Lin Zhuo [a], Yingjie Xiao [a,c], Jianbing Yan [a,c], Wenyu Yang [a,b,*]

[a] National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan, Hubei 430070, China

[b] College of Informatics, Huazhong Agricultural University, Wuhan, Hubei 430070, China

[c] Hubei Hongshan Laboratory, Wuhan, Hubei 430070, China

[1] These authors contributed equally.

[*] Correspondence author

E-mail address: yangwenyu@mail.hzau.edu.cn (W. Y.).

Abstract

Phenotypic prediction is a promising strategy for accelerating plant breeding. Data from multiple sources (called multi-view data) can provide complementary information to characterize a biological object from various aspects. By integrating multi-view information into phenotypic prediction, a multi-view best linear unbiased prediction (MVBLUP) method was proposed in this paper. To measure the importance of multiple data views, the differential evolution algorithm with an early stopping mechanism was used, by which we obtained a multi-view kinship matrix and then incorporated it into the BLUP model for phenotypic prediction. To further illustrate the characteristics of MVBLUP, we performed the empirical experiments on four multi-view datasets in different crops. Compared to the single-view method, the prediction accuracy of the MVBLUP method has improved by 0.038 to 0.201 on average. The results demonstrate that the MVBLUP is an effective integrative prediction method for multi-view data.

Keywords

Multi-view data, Best linear unbiased prediction, Similarity function, Phenotype prediction, Differential evolution algorithm

Introduction

Phenotype prediction is a powerful tool that allows for the early assessment of traits in individuals prior to planting, thereby accelerating the breeding process and significantly reducing its duration. This prediction is primarily achieved through genomic prediction (GP), a concept initially introduced by Meuwissen for animal breeding (Meuwissen et al., 2001). Since then, various methods have been employed to predict traits, broadly categorized into traditional statistical methods and machine learning approaches. Traditional statistical methods encompass best linear unbiased predictions (BLUP) (Henderson, 1975), least absolute shrinkage and selection operator (LASSO) (Usai et al., 2009), and Bayesian-based methods such as Bayes A, Bayes B, Bayesian LASSO (Meuwissen et al., 2001; Yi and Xu, 2008; de los Campos et al., 2009). On the other hand, machine learning approaches include support vector machine (SVM) (Maenhout et al., 2007), random forest (RF) (Holliday et al., 2012), deep convolutional neural networks (DeepGS) (Ma et al., 2018), and deep neural network for genomic prediction using multi-omics data (DNNGP) (Wang et al., 2023). Notably, DNNGP incorporates a batch normalization (BN) layer to mitigate overfitting and can be viewed as an advanced version of DeepGS. Other notable approaches include extreme gradient boosting (XGBoost) (Chen and Guestrin, 2016; Xu et al., 2016) and its faster version, light gradient boosting machine (LightGBM) (Yan et al., 2021).

Multi-view data refers to data from multiple sources that offer complementary information to characterize a biological object from various perspectives. These data can include different groups of samples measured by the same feature set (multi-class data), the same samples with various feature sets (e.g., multi-omics data), the same samples by the same set of features under different conditions (e.g., multi-environmental data), or different features and different samples in the same system (multi-relational data) (Li et al., 2018). Such data are ubiquitous in the real world. For example, a sample can be characterized by its genotype, gene expression levels, and metabolomic data, each serving as a unique view of the sample. Compared to single-view data, multi-view data

can provide more complementary information, and thus, effective integration of multi-view data has the potential to enhance model prediction performance (Serra et al., 2015; Dimitrakopoulos et al., 2017; Montesinos-López et al., 2018).

Recent studies have demonstrated that integrating multi-view data can lead to higher prediction accuracy. For example, incorporating transcript levels from seedlings with genetic markers into a joint model improved the prediction of mature maize traits (Azodi et al., 2019). By integrating metabolomics data into genomic prediction of hybrid yield of rice, predictability was enhanced by approximately 30% (Xu et al., 2016). During the learning process of a LASSO model with genome-wide markers, sequential integration of transcriptome and metabolome features allowed for iterative learning of three layers of features, resulting in significant improvement in rice yield trait prediction (Hu et al., 2019). Additionally, incorporating genotype × environment interaction (G×E) into a GP model has been shown to improve prediction accuracy (Montesinos-López et al., 2018; Crossa et al., 2021; Xu et al., 2022; Barreto et al., 2024).

Despite the demonstrated efficiency of multi-view integration prediction, there is still considerable room for improvement in learning from different data views. The relationship among different views is often complex, with different data sources potentially containing varying amounts of information and noise. The quality of data typically varies across different samples, meaning that one view may be informative for one sample but not for another. Existing multi-view methods often treat each view with equal importance, tune their weights to fixed values, or integrate them with a black-box machine learning framework (Wang et al., 2021). Therefore, there is a need to develop new methods for integrating multi-view information.

The differential evolution algorithm (DE), first proposed by Storn and Price, is a population-based evolutionary algorithm (EA) designed to search for a parameter set that maximizes a target function (called a fitness function) (Storn and Price, 1997). The algorithm mimics natural evolution through an iteration process involving mutation, crossover, and selection, evolving the population towards better solutions. In the mutation phase, DE generates a new individual (called the mutant vector) by computing the difference between two randomly selected individuals, scaling the difference by a factor, and adding the result to a third randomly selected individual. At the crossover stage, a trial individual can be generated by combining the mutant individual with a target individual with a certain probability determining which genes come from the mutant and which from the target individual. The selection step evaluates the fitness value of the trial

individual and compares this fitness value with that of the target individual. If the fitness of the trial individual is better, replace the target individual with the trial individual in the population. Otherwise, keep the target individual. It ensures that only individuals with improved fitness are retained in the population.

In this study, we adapted DE to establish an adaptive multi-view integration strategy to better measure the importance of different views. By combining this adaptive integration strategy with the common statistical prediction model, BLUP, we proposed a multi-view best linear unbiased prediction (MVBLUP) method for phenotype prediction. The schematic workflow of the method is illustrated in Fig. 1, and details of the algorithm can be found in the Method subsection of the paper. To evaluate the performance of the MVBLUP, we compared its prediction accuracy with BLUP, LASSO, and XGBoost using single-view and multi-view data from tomato, rice, and maize datasets of diverse sizes. Numerical results demonstrate that MVBLUP is a promising and practical approach for integrating multi-view data for phenotype prediction.

Results

A comparative analysis of the results was conducted by using MVBLUP, BLUP, LASSO, and XGBoost, each evaluated separately on the datasets Tomato332, Rice210, Maize368, and Maize282. The standard for assessing these comparison results was the average prediction accuracy derived from 50 random five-fold cross-validation. Specifically, this involved calculating the mean Pearson's correlation coefficient (PCC) between the predicted and observed values across 50 random tests.

MVBLUP for predicting solid content trait (SSC) trait of Tomato332

The analysis was initiated by assessing the prediction accuracy of MVBLUP for the fruit-soluble SSC trait of Tomato332. Three distinct views, single nucleotide polymorphisms (SNP), insertions and deletions (InDel), and structural variants (SV), were selected to construct a multi-view prediction.

To highlight the distinct information conveyed by different views, heatmaps of their respective kinship matrices were presented in Fig. 2A. These heatmaps unveiled discernible

variations in the similarity patterns across the data from the three different views. For each view, LASSO, BLUP, and XGBoost were employed independently to predict the SSC trait.

The results demonstrated that prediction accuracy varied depending on the view data utilized, even when the same method was applied. This further illustrated the distinctiveness of the information carried by different view data (Fig. 2B). Notably, the BLUP method exhibited the highest prediction accuracy across all three types of view data. Given the robust performance of BLUP (Xu et al., 2024), multi-view information was integrated based on BLUP, leading to the development of the MVBLUP method.

Comparisons of BLUP using single-view data, MVBLUP with pairwise integrative views, and MVBLUP with all three integrative views were given (Fig. 2C). Results revealed that integrating additional views improved prediction accuracy. BLUP with a single InDel feature performed the worst (0.347). When MVBLUP incorporated pairwise-view features, the prediction accuracy rose to 0.384. With all three views integrated, MVBLUP achieved the highest accuracy of 0.398.

Including both SNP and InDel features in LASSO and XGBoost models (Fig. 2D) led to prediction accuracies of 0.200 and 0.319, respectively, making improvements over single-view data. This trend aligns with the results observed in the MVBLUP model (Fig. 2C), suggesting that MVBLUP's enhanced prediction accuracy is partly attributed to the complementarity of multi-view data.

It is crucial to note that directly integrating multi-view information into a model can sometimes reduce prediction accuracy if the information is redundant or incompatible. For instance, when LASSO and XGBoost methods incorporated the SV feature, their prediction accuracy decreased by 0.218 and 0.163 respectively, compared to when they used SNP feature (Fig. 2B). Additionally, the introduction of the SV feature further reduced the prediction accuracies of both methods to 0.131 and 0.261 respectively (Fig. 2D). This highlights the significance of extracting complementary information while excluding redundant and incompatible information during the process of multi-view data fusion.

One solution to this challenge is assigning weights to the multi-view data entering the model. Based on BLUP, three weight assignment methods: uniform weights, weights based on the prediction accuracy of the training set, and optimal weights learned by a differential evolution algorithm (MVBLUP), were employed for multi-view integration. Comparative analysis revealed

that MVBLUP outperformed the other two methods, achieving a 0.022 improvement in prediction accuracy (Fig. 2D). Consequently, we adopted the MVBLUP approach for integrating multi-view data.

MVBLUP for predicting four traits of Rice210

For the dataset Rice210, three distinct types of omics data, genomic (G), gene expression (E), and metabolomic (M), were meticulously selected and seamlessly integrated into the MVBLUP framework to predict four traits: grain number per panicle, 1000 grain weight, yield per plant, and tiller number per plant. Notably, MVBLUP demonstrated remarkable prediction accuracy, outperforming the single-view BLUP method for three out of the four traits being assessed (Fig. 3A–3C). Specifically, MVBLUP surpassed the single-view BLUP method by approximately 0.05 in predicting grain number per panicle (Fig. 3A). For the yield per plant trait, MVBLUP achieved an impressive prediction accuracy of 0.724, which has been significantly improved by 0.296 compared to the single-view BLUP using genomic data alone (Fig. 3C).

However, it is worth mentioning that the results for the tiller number per plant trait were somewhat unexpected. In this case, MVBLUP's accuracy of 0.707 was slightly lower than that of the single-view BLUP with genomic data which was 0.713 (Fig. 3D).

MVBLUP for predicting eight traits of Maize368

In the context of the Maize368 dataset, MVBLUP was employed to predict eight diverse traits: heading date, silking time, pollen shedding, cob diameter, ear diameter, ear length, ear leaf width, ear leaf length. This prediction was facilitated by the integration of three-view data, comprising genomic (G), gene expression (E), and metabolomic (M). MVBLUP demonstrated remarkable prediction accuracies for seven of these traits (Fig. 4A–4G). Specifically, for the trait of heading date, MVBLUP surpassed the prediction accuracy of the single-view BLUP method with genomic data by a margin of approximately 0.041 (Fig. 4A). In the case of silking time, MVBLUP achieved an accuracy of 0.632, marking a 0.073 increase in accuracy compared to the single-view BLUP utilizing genomic data (Fig. 4B). The only trait where MVBLUP's performance was somewhat

less impressive was ear leaf length, with a prediction accuracy of 0.609, which was marginally lower than the accuracy achieved by the single-view BLUP with genomic data (Fig. 4H).

MVBLUP for predicting six traits of Maize282

For the Maize282 dataset, MVBLUP was utilized to predict six traits: days to anthesis, plant height, ear height, node number below ear, leaf width, weight of 20 kernels. This prediction was enabled by the integration of eight views, including genomic data (G) alongside gene expression data sourced from seven distinct tissues (E1–E7). MVBLUP exhibited superior prediction accuracy for five out of the six traits when compared to BLUP utilizing individual views (Fig. 5A–5E). Notably, MVBLUP surpassed the prediction accuracy of the single-view BLUP method with genomic data for the trait of node number below ears by a margin of approximately 0.039 (Fig. 5D). For the trait of weight of 20 kernels (Fig. 5F), while MVBLUP's accuracy of 0.495 was slightly lower than the optimal accuracy of 0.506 achieved by BLUP with the E3 view (gene expression data from the tissue of the third leaf from the base), it still exceeded the performance of the single-view BLUP method, which utilized genomic data, by a notable margin of approximately 0.027.

The efficiency of MVBLUP in the Maize282 dataset was determined by the optimal weight calculated by the DE algorithm. To demonstrate the convergence of the DE algorithm, the iteration process of MVBLUP on Maize282 was shown (Fig. 6). For this dataset, the initial population size was set as 40, with eight weight parameters needing optimization. It showed that the algorithm converged steadily to the optimal solution as the iteration progressed (Fig. 6). Although the maximum iteration number of the DE algorithm was set at 50, the algorithm stabilized after 43 iterations. The early stopping mechanism, which was triggered when the maximum error of the cost function value at two adjacent iteration points fell below a specified tolerance, was instrumental in saving computational costs.

To further validate the effectiveness of the MVBLUP method, numerical experiments analogous to those performed on the Tomato332 dataset were conducted using the Rice210, Maize368, and Maize282 datasets. These experiments involved comparisons with LASSO, XGBoost, BLUP integrating each view with uniform weights (BLUP_W1), and BLUP integrating each view with fixed weights determined by the average accuracy of five-fold cross-validation on single-view data training sets (BLUP_W2). The results of these comparisons are presented in Figs.

S1–S3, which showed that MVBLUP outperformed the other methods in most cases. Additionally, the average running time of MVBLUP was significantly faster than the XGBoost method on the Rice210, Maize368, and Maize282 datasets, with average running times of 28.3 seconds, 68.9 seconds, and 283.8 seconds, respectively (Table S1). However, it is worth noting that MVBLUP required the longest average running time on the Tomato332 dataset, at 58.5 seconds.

Discussion

In this study, we aimed to enhance phenotype prediction by integrating multiple data views through the application of the MVBLUP method. Compared with existing multi-view integrating methods that assign equal importance to each view, MVBLUP offers greater interpretability by an adaptively adjusting weights strategy adjusted with the DE algorithm to quantify the contribution of different views. The strengths of MVBLUP lie in the complementary information derived from multi-view data, the accuracy of single-view models, and the synergistic integration of various views via weights learned by the DE algorithm. These advantages may vary across populations, traits, and datasets.

Recently, a GA-BLUP method, which combines BLUP with a genetic algorithm (GA) to select trait-related markers, has emerged as a highly precise genomic prediction method, particularly for traits with low heritability (Xu et al., 2024). MVBLUP shares some similarities with GA-GBLUP in that both utilize evolutionary algorithms for selection. Specifically, MVBLUP employs DE for selecting multi-view data, whereas GA-GBLUP uses GA for marker selection. Although DE and GA share fundamental operators like mutation, crossover, and selection, they differ in key aspects. Notably, the mutation operator in DE fundamentally differs from that in GA. In GA, mutation typically involves randomly altering individual bits or genes in a chromosome, whereas DE employs a differential mutation strategy that explores the search space more efficiently and effectively. In addition, GA excels in discrete optimization problems due to its encoding and decoding mechanism tailored for combinatorial optimization, whereas DE performs better in continuous optimization problems, which are more suitable for integrating multi-view data in this study.

MVBLUP demonstrated superior prediction accuracy compared to BLUP, LASSO, and XGBoost, both with single-view and multi-view data, across four datasets, highlighting its

effectiveness and applicability. However, it is worth noting that MVBLUP slightly underperformed in three specific cases. For instance, in predicting the tiller number per plant trait of the Rice210 dataset, the accuracy of MVBLUP (0.707) was marginally lower than that of single-view BLUP with genomic data (0.713). We hypothesize that the slight decrement in performance for certain traits could be attributed to discrepancies in data distributions between the training and test datasets, which may lead to the model excelling on the training set but struggling with generalization to the test set. During MVBLUP's learning process, optimal weights were selected based on the highest fitness function value on the training set. However, in some instances, despite achieving optimized training performance, the testing accuracy fell short of expectations. To illustrate this phenomenon, we randomly divided the Rice210 dataset into training and test sets 50 times and calculated the accuracy using MVBLUP and single-view BLUP based on genomic data after each division. MVBLUP exhibited better training accuracy than the single-view BLUP method in each training set (Fig. S4A). However, in terms of test performance, MVBLUP showed lower prediction accuracy in some test sets (Fig. S4B). This comparison underscores the potential discrepancy between training performance and actual testing outcomes, which could be partially mitigated by increasing the sample size of the dataset.

Moreover, a deeper exploration of MVBLUP's enhancements is crucial for future research. In this study, we assigned weights using the DE algorithm, commonly used in vast and complex parameter spaces. However, with a maximum of eight multi-view data sources, the full advantages of DE were not fully demonstrated. Importantly, our MVBLUP framework is inherently scalable to accommodate a broader range of multi-view scenarios. In breeding, MVBLUP offers a solution for efficiently integrating multi-view data including genomic data, red-green-blue (RGB) images, and spectrum image-based phenomic data. The use of Unmanned Aerial Vehicles (UAVs) for high-throughput phenotyping will drastically accelerate the collection of multi-view data related to plant physiological status throughout the growth period, achieving both efficiency and cost-effectiveness. Alternatively, if focus solely on genomic data, we can categorize the data by chromosomes and assign weights accordingly. Furthermore, we can assign weights to each SNP marker and employ the DE method to learn and optimize these weights. As an evolutionary algorithm, DE has the potential to avoid local optima but may compromise computational efficiency. Therefore, future research could explore efficient alternatives, such as improved DE strategies with accelerated convergence (Bilal et al., 2020).

Materials and methods

Data sources

The Tomato332 dataset comprises 332 materials from three tomato subspecies: currant tomato (PIM), cherry tomato (CER), and large-fruited cultivated tomato (BIG). The genotype data for this dataset is a call set named TGG1.1-332 from the tomato graph pangenome, which encompasses 6,971,059 SNPs, 657,549 InDels, and 54,838 SVs. A crucial phenotypic trait in this dataset is the fruit soluble solids content (SSC), which is significant for both yield and flavor (Zhou et al., 2022). By applying Principal Component Analysis (PCA) to the genotype data, the dataset was reduced to 220 SNPs, 289 InDels, and 277 SVs, which were then used as multi-view data for predicting the SSC trait (Wang et al., 2023).

The Rice210 dataset consists of 210 recombinant inbred lines (RILs), obtained through the crossing two rice varieties Zhenshan 97 and Minghui 63 (Hua et al., 2003). Sequencing of these RILs resulted in the identification of 270,820 high-quality SNPs and 1,619 bins based on recombination breakpoints, serving as the genotype data (Yu et al., 2011). Ribonucleic Acid (RNA) was extracted from the flag leaves of the RILs during the heading stage between 8:00 and 9:30 AM, and the expression levels of 24,994 genes were quantified using a microarray sequencing platform, providing transcriptomic data (Wang et al., 2014). Metabolomic data included 1,000 metabolites sourced from two tissues: flag leaves at the heading stage and seeds 72 hours post-germination (Gong et al., 2013). Four key agronomic traits—yield per plant, tiller number per plant, grain number per panicle, and 1000-grain weight—were collected in 2008 and 2009 from a field experiment conducted at the Farm of Huazhong Agricultural University in Wuhan, China (Yu et al., 2011).

The Maize368 dataset includes 368 maize inbred lines derived from a broadly representative association mapping population encompassing tropical, subtropical, and temperate germplasm (Yang et al., 2011). These inbred lines were genotyped using four different genotyping platforms, resulting in the identification of 1.25 million high-quality SNP markers as the genotype data (Liu et al., 2017). RNA was extracted from immature kernels at 15 days post-pollination and sequenced to obtain expression levels for 28,769 genes (Fu et al., 2013). Additionally, metabolic profiling of

mature maize kernels led to the identification of 749 non-targeted metabolites (Wen et al., 2014). This study utilized eight agronomic traits previously analyzed in a Genome-Wide Association Studies (GWAS) study (Yang et al., 2014), including cob diameter, ear diameter, ear leaf width, ear length, ear leaf length, heading date, pollen shedding and silking time. These traits were collected across five environments, and their average values were used for phenotypic prediction (Yang et al., 2014).

The Maize282 dataset comprises 282 maize inbred lines sourced from a US maize association mapping panel (Flint-Garcia et al., 2005). Genotyping of these inbred lines was conducted using the Illumina MaizeSNP50 BeadChip, resulting in the identification of 50,878 high-quality SNP markers as the genotype data (Ganal et al., 2011). RNA extraction and sequencing were performed on seven different tissues at specified times and locations, including the base and tip of the third leaf collected between 10:30 and 12:00, the root of a 2 cm germinated seedling, and the entire shoot of the germinated seedling collected between 11:00 and 13:00 on the day of germination, developing kernels post-pollination collected between 11:00 and 13:00, and mature leaf samples collected from a 1 cm section adjacent to the midrib of the second leaf below the tassel between 11:00 and 13:00, and also between 23:00 and 1:00 (Kremling et al., 2018). Quantitative analysis of messenger RNA (mRNA) expression levels provided transcriptomic data. Six important agronomic traits, including plant height, weight of 20 kernels, node number below ear, days to anthesis, ear height, and leaf width, were used in this study (Flint-Garcia et al., 2005).

Method

MVBLUP is a prediction model that builds upon the BLUP framework and incorporates the DE algorithm to dynamically determine the optimal integrating weight of multi-view features. To ensure comprehensiveness, we first introduce the principle of BLUP and DE.

Best linear unbiased prediction

The BLUP approach relies on a mixed linear model. The fundamental model could be articulated as:

$$y = X\beta + \xi + \varepsilon,$$

where: y is an $n \times 1$ vector of phenotypic values of a quantitative trait for n individuals, X is an $n \times p$ design matrix, β is a $p \times 1$ vector of fixed effects, $\xi \sim N(0, K\sigma^2)$ is a $n \times 1$ vector of random effects, $\varepsilon \sim N(0, I\sigma_e^2)$ is a $n \times 1$ vector of residual errors, K represents the relationship

between individuals, σ^2 and σ_e^2 are variance associated with random effects and residual errors respectively, I is an identity matrix. The random effects vector ξ can be derived using:

$$\xi = \sigma^2 K V^{-1}(y - X\hat{\beta}),$$

where $V = K\sigma^2 + I\sigma_e^2$ *and* $\hat{\beta} = (X^T V^{-1} X)^{-1}(X^T V^{-1} y)$.

The similarity function and kinship matrix

Let x_i denotes the input feature vector of the *i*th individual (for example, the expression level of the *i*th individual). The similarity between the *i*th and the *j*th individual is defined as:

$$k(x_i, x_j) = \frac{(x_i, x_j)}{\sqrt{(x_i, x_i)}\sqrt{(x_j, x_j)}},$$

where (x_i, x_j) represents the inner product of the vector x_i and x_j.

Based on the similarity function, we define a single view initial kinship matrix A_0 as $A_0 = \begin{pmatrix} d_{11} & \cdots & d_{1n} \\ \vdots & \ddots & \vdots \\ d_{n1} & \cdots & d_{nn} \end{pmatrix}$, where $d_{ij} = (x_i, x_j)$ and n is the number of individuals. Further, we can define the single view kinship matrix A as $A = \begin{pmatrix} k_{11} & \cdots & k_{1n} \\ \vdots & \ddots & \vdots \\ k_{n1} & \cdots & k_{nn} \end{pmatrix}$, where $k_{ij} = k(x_i, x_j)$. Finally, we define the relationship matrix K which integrates multiple views of data as: $K = w_1^2 * A_{10} + w_2^2 * A_{20} + \cdots + w_m^2 * A_{m0}$, where A_{i0} is the *i*th view initial kinship matrix, m is the number of view and w_j is the weight of the *j*th view data. Here the new multi-view relationship matrix can be seen as a generalization of single view kinship matrix.

Differential evolution algorithm

In this study, DE is employed to identify the optimal weights for each view's initial kinship matrix. The DE process encompasses initialization, mutation, crossover, and selection. The steps are detailed as follows:

Step 1. Initialization. An initial population $P_0 = \{X_i^0 : i = 1,2, \ldots, n\}$ is generated as follows:

$$X_i^0 = X_{low} + (X_{upp} - X_{low}) * rand(0,1),$$

where n is the size of population, and X_{low} and X_{upp} are lower and upper bounds of search space, respectively.

Step 2. Mutation. The ith mutant individual in the gth vector generation is created according to

the following:

$$V_i^g = X_{r1}^g + c(X_{r2}^g - X_{r3}^g),$$

where X_{r1}^g, X_{r2}^g and X_{r3}^g are the randomly selected individuals from the parent population, $r1$, $r2$, $r3 \in \{1,2,\dots,n\}$, and $c \in (0,1)$ is the scaling factor.

Step 3. Crossover. The ith crossover individual in the gth trial vector is generated as follows,

$$U_i^g = (U_{1i,}^g U_{2i,}^g \dots, U_{mi,}^g),$$

$$U_{ji}^g = \begin{cases} V_{ji}^g, if\ randb(j) \le \delta\ or\ j = rnbr(i) \\ X_{ji}^g, if\ randb(j) > \delta\ and\ j = rnbr(i) \end{cases}, j = 1,2,\cdots,m,$$

where $randb(j) \in [0,1]$ is a uniform random number, $\delta \in [0,1]$ is a predefined crossover parameter, $ranbr(i) \in \{1,2,\mathrm{L}\ ,m\}$ is an index selected randomly.

Step 4. Selection. Determine whether the trial vector in the crossover step should be included in next generation by a strategy as follows,

$$X_i^{k+1} = \begin{cases} U_i^k, & \text{if } f(U_i^k) > f(X_i^k) \\ X_i^{k+1}, & \text{else.} \end{cases}$$

where $f(X)$ is a fitness function.

Step 5. Stopping criterion. If maximum error of the fitness function value between X_i^{k+1} and X_i^k is less than tolerance ε (i.e., an early stopping mechanism), or the maximum iteration count K is reached, then DE algorithm will be stopped.

Prior to training, five parameters must be set: population size n, scaling factor c, crossover parameter δ, tolerance ε and maximum iteration parameter K. In our experiment, population size n is set as five times the number of feature views. Specifically, $n = 15$ for Tomato332, Rice210 and Maize368, $n = 40$ for Maize282. Both scaling factor c and crossover parameter δ are set as 0.5, tolerance $\varepsilon = 0.0001$ and maximum iteration number $K = 50$.

The multi-view best linear unbiased prediction procedure

MVBLUP is a flexible machine learning algorithm that integrates adaptively multi-view data for phenotype prediction. The MVBLUP process involves:

Step 1 Normalization. Input vectors are normalized using the Z-score method, and initial weights are assigned randomly to establish the initial population $P_0 = \{W_i^0 : i = 1,2,\dots,n\}$ for DE.

Step 2 Training. Repeat the following steps *K* times or stop with an early stopping mechanism:

Step 2.1. Set $k = 1$. By utilizing the "Mutation and Crossover" steps within the DE algorithm, the n weight vectors are renewed. Subsequently, these updated weight vectors are employed to compute n multi-view kinship matrices through the application of the similarity function.

Step 2.2. Using "Selection" steps in the DE algorithm to renew the weight vector, we obtain the initial population of the next generation. The "Selection" step utilizes the average prediction accuracy derived from five-fold cross-validation on the training set as the fitness function, which can be mathematically represented as $f(W) = E(\frac{(y - E(y), \hat{y} - E(\hat{y}))}{\|y - E(y)\|\|\hat{y} - E(\hat{y})\|})$, where E represents expectation, y is the observed value vector for a particular trait, $\hat{y}$ is the predicted value vector for the corresponding trait, and $\|\cdot\|$ represents the 2-norm, $\|y\| = \sqrt{(y, y)}$.

Step 2.3 (Stopping criterion). If the difference between $f(W^{k+1})$ and $f(W^k)$ is less than tolerance, then optimal weight vector is found, training will be stopped; Else, $k = k + 1$, go to Step 2.1.

Step 3 (Prediction). Phenotypes are predicted using the optimized multi-view kinship matrix.

Methods used for comparisons

The predictive capabilities of MVBLUP were initially compared with BLUP when utilizing single-view data. Subsequently, two additional prevalent techniques: least absolute shrinkage and selection operator (LASSO) and extreme gradient boosting (XGBoost), were applied to both single-view and multi-view datasets for comparative purposes.

LASSO aims to identify an optimal linear model represented as $y = X\alpha + \varepsilon$, where X is an $n \times d$ matrix , y is a $n -$ dimensional vector, ε denotes noise and α serves as the weight vector. To determine an appropriate α, LASSO can be formulated as the following optimization problem:

$$min_{\alpha} \frac{1}{2} \sum_{i=1}^{n} (X_i \alpha - y_i)^2 + \sigma \|\alpha\|_1,$$

with σ being a regularization parameter. In this study, the LASSO method was executed using the GLMNET/R package (Friedman et al., 2010).

XGBoost method, on the other hand, focuses on constructing an ensemble of trees, utilizing *K* additive functions to predict the output,

$$\hat{y}_i = \varphi(X_i) = \sum_{k=1}^{K} f_k(X_i),\ f_k \in F$$

where $X_i \in R^d, y_i \in R, F = \{f(x) = \omega_{q(x)}\}(q:\ R^d \to T, \omega \in R^T)$ represents the space of regression trees (also called as CART), q represents the structure of each tree, T is the number of leaves in the tree, each f_k corresponds to an independent tree structure q and leaf weight ω.

To learn the functions in the model, XGBoost can be framed as the following optimization problem,

$$L(\varphi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k),$$

where l is a differentiable convex loss function measuring the discrepancy between the prediction $\hat{y}_i$ and the target y_i, $\Omega(f_k) = \gamma T + \frac{\sigma}{2}\|\omega\|^2$ serves as a regularization term penalizing model complexity, γ and σ are regularization parameters (Chen and Guestrin, 2016). In this study, XGBoost was implemented using the xgboost/R package.

Data availability

The demo data, R scripts, and tutorial of MVBLUP algorithm are available for public access on GitHub at the following link: https://github.com/bjwu555/MVBLUP.

CrediT authorship contribution statement

Bingjie Wu: Methodology, Data curation, Investigation, Validation, Writing - Original draft. **Huijuan Xiong**: Investigation, Validation, Writing - Original draft, Writing - Review & Editing. **Lin Zhuo**: Data curation, Investigation. **Yingjie Xiao**: Conceptualization, Project administration, Resources, Supervision. **Jianbing Yan**: Conceptualization, Project administration, Resources, Supervision. **Wenyu Yang**: Conceptualization, Methodology, Validation, Project administration, Supervision, Writing - Review & Editing.

Conflict of interest

The authors declare that they have no competing interests.

Acknowledgments

This work was supported by National Natural Science Foundation of China (32122066, 32201855), and STI2030—Major Projects (2023ZD04076).

References

Azodi, C.B., Jeremy, P., Robert, V.B., Gustavo, D.L.C.,Shin-Han, S., 2019. Transcriptome-based prediction of complex traits in maize. Plant Cell 32, 139-151.

Barreto, C.A.V., Dias, K.O.D.G., Sousa, I.C.D., Azevedo, C.F., Nascimento, A.C.C., Guimares, L.J.M., Guimares, C.T., Pastina, M.M.,Nascimento, M., 2024. Genomic prediction in multi-environment trials in maize using statistical and machine learning methods. Sci. Rep. 14, 1062.

Bilal, Pant, M., Zaheer, H., Garcia-Hernandez, L.,Abraham, A., 2020. Differential evolution: A review of more than two decades of research. Eng. Appl. Artif. Intel. 90, 103479.

Chen, T.,Guestrin, C. 2016. Xgboost: A scalable tree boosting system. Paper presented at: Proc. 22nd ACM SIGKDD Inter. Conf. KDDM Association for Computing Machinery, San Francisco California USA.

Crossa, J., Fritsche-Neto, R., Montesinos-Lopez, O.A., Costa-Neto, G., Dreisigacker, S., Montesinos-Lopez, A.,Bentley, A.R., 2021. The modern plant breeding triangle: Optimizing the use of genomics, phenomics, and enviromics data. Front. Plant Sci. 12, 651480.

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A.s., Manfredi, E., Weigel, K.,Cotes, J.M., 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182, 375-385.

Dimitrakopoulos, L., Prassas, I., Diamandis, E.P.,Charames, G.S., 2017. Onco-proteogenomics: Multi-omics level data integration for accurate phenotype prediction. Crit. Rev. Clin. Lab. Sci. 54, 414-432.

Flint-Garcia, S.A., Thuillet, A.-C., Yu, J., Pressoir, G., Romero, S.M., Mitchell, S.E., Doebley, J., Kresovich, S., Goodman, M.M.,Buckler, E.S., 2005. Maize association population: A high-resolution platform for quantitative trait locus dissection. Plant J. 44, 1054-1064.

Friedman, J.H., Hastie, T.,Tibshirani, R., 2010. Regularized paths for generalized linear models via coordinate descent. J. Stat. Softw. 33, 1-22.

Fu, J., Cheng, Y., Linghu, J., Yang, X., Kang, L., Zhang, Z., Zhang, J., He, C., Du, X., Peng, Z., *et al.*, 2013. Rna sequencing reveals the complex regulatory network in the maize kernel. Nat. Commun. 4, 2832.

Ganal, M.W., Durstewitz, G., Polley, A., Bérard, A., Buckler, E.S., Charcosset, A., Clarke, J.D., Graner, E.-M., Hansen, M., Joets, J., *et al.*, 2011. A large maize (zea mays l.) snp genotyping array:

Development and germplasm genotyping, and genetic mapping to compare with the b73 reference genome. PLoS One 6, e28334.

Gong, L., Chen, W., Gao, Y., Liu, X., Zhang, H., Xu, C., Yu, S., Zhang, Q.,Luo, J., 2013. Genetic analysis of the metabolome exemplified using a rice population. Proc. Natl. Acad. Sci. U.S.A. 110, 20320-20325.

Henderson, C.R., 1975. Best linear unbiased estimation and prediction under a selection model. Biom. 31, 423-447.

Holliday, J.A., Wang, T.,Aitken, S., 2012. Predicting adaptive phenotypes from multilocus genotypes in sitka spruce (picea sitchensis) using random forest. G3:Genes Genom. Genet. 2, 1085-1093.

Hu, X., Xie, W., Wu, C.,Xu, S., 2019. A directed learning strategy integrating multiple omic data improves genomic prediction. Plant Biotechnol. J. 17, 2011-2020.

Hua, J., Xing, Y., Wu, W., Xu, C., Sun, X., Yu, S.,Zhang, Q., 2003. Single-locus heterotic effects and dominance by dominance interactions can adequately explain the genetic basis of heterosis in an elite rice hybrid. Proc. Natl. Acad. Sci. U.S.A. 100, 2574-2579.

Kremling, K.A.G., Chen, S.Y., Su, M.H., Lepak, N.K., Romay, M.C., Swarts, K.L., Lu, F., Lorant, A., Bradbury, P.J.,Buckler, E.S., 2018. Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. Nature 555, 520-523.

Li, B., Zhang, N., Wang, Y.-G., George, A.W., Reverter, A.,Li, Y., 2018. Genomic prediction of breeding values using a subset of snps identified by three machine learning methods. Front. Genet. 9, 237.

Liu, H., Luo, X., Niu, L., Xiao, Y., Chen, L., Liu, J., Wang, X., Jin, M., Li, W.,Zhang, Q., 2017. Distant eqtls and non-coding sequences play critical roles in regulating gene expression and quantitative trait variation in maize. Mol. Plant 10, 414-426.

Ma, W., Qiu, Z., Song, J., Li, J., Cheng, Q., Zhai, J.,Ma, C., 2018. A deep convolutional neural network approach for predicting phenotypes from genotypes. Planta 248, 1307-1318.

Maenhout, S., Baets, B.D., Haesaert, G.,Bockstaele, E.V., 2007. Support vector machine regression for the prediction of maize hybrid performance. Theor. Appl. Genet. 115, 1003-1013.

Meuwissen, T.H., Hayes, B.J.,Goddard, M.E., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819-1829.

Montesinos-López, O.A., Montesinos-López, A., Crossa, J., Gianola, D., Hernández-Suárez, C.M.,Martín-Vallejo, J., 2018. Multi-trait, multi-environment deep learning modeling for genomic-enabled prediction of plant traits. G3:Genes Genom. Genet. 8, 3829-3840.

Serra, A., Fratello, M., Fortino, V., Raiconi, G., Tagliaferri, R.,Greco, D., 2015. Mvda: A multi-view genomic data integration methodology. BMC Bioinf. 16, 261.

519 Storn, R.,Price, K., 1997. Differential evolution – a simple and efficient heuristic for global optimization
520 over continuous spaces. J. Global Optim. 11, 341-359.
521 Usai, M.G., Goddard, M.E.,Hayes, B.J., 2009. Lasso with cross-validation for genomic selection. Genet.
522 Res. 91, 427-436.
523 Wang, J., Yu, H., Weng, X., Xie, W., Xu, C.-g., Li, X., Xiao, J.,Zhang, Q., 2014. An expression quantitative
524 trait loci-guided co-expression analysis for constructing regulatory network using a rice
525 recombinant inbred line population. J. Exp. Bot. 65, 1069 - 1079.
526 Wang, K., Abid, M.A., Rasheed, A., Crossa, J., Hearne, S.,Li, H., 2023. Dnngp, a deep neural network-
527 based method for genomic prediction using multi-omics data in plants. Mol. Plant 16, 279-293.
528 Wang, T., Shao, W., Huang, Z., Tang, H.,Huang, K., 2021. Mogonet integrates multi-omics data using
529 graph convolutional networks allowing patient classification and biomarker identification. Nat.
530 Commun. 12, 3445.
531 Wen, W., Li, D., Li, X., Gao, Y., Li, W., Li, H., Liu, J., Liu, H., Chen, W.,Luo, J., 2014. Metabolome-based
532 genome-wide association study of maize kernel leads to novel biochemical insights. Nat. Commun.
533 5, 3438.
534 Xu, S., Xu, Y., Gong, L.,Zhang, Q., 2016. Metabolomic prediction of yield in hybrid rice. Plant J. 88, 219-
535 227.
536 Xu, Y., Zhang, X., Li, H., Zheng, H., Zhang, J., Olsen, M.S., Varshnev, R.K., M.Prasanna, B.,Qian, Q.,
537 2022. Smart breeding driven by big data,artificial intelligence,and integrated genomic-enviromic
538 prediction. Mol. Plant 15, 1664-1695.
539 Xu, Y., Zhang, Y., Cui, Y., Zhou, K., Yu, G., Yang, W., Wang, X., Li, F., Guan, X., Zhang, X., *et al.*, 2024.
540 Ga-gblup: Leveraging the genetic algorithm to improve the predictability of genomic selection.
541 Briefings Bioinf. 25, bbae385.
542 Yan, J., Xu, Y., Cheng, Q., Jiang, S., Wang, Q., Xiao, Y., Ma, C., Yan, J.,Wang, X., 2021. Lightgbm:
543 Accelerated genomically designed crop breeding through ensemble learning. Genome Biol. 22, 271.
544 Yang, N., Lu, Y., Yang, X., Huang, J., Zhou, Y., Ali, F., Wen, W., Liu, J., Li, J.,Yan, J., 2014. Genome
545 wide association studies using a new nonparametric model reveal the genetic architecture of 17
546 agronomic traits in an enlarged maize association panel. PLos Genet. 10, e1004573.
547 Yang, X., Gao, S., Xu, S., Zhang, Z., Prasanna, B.M., Li, L., Li, J.,Yan, J., 2011. Characterization of a
548 global germplasm collection and its potential utilization for analysis of complex quantitative traits
549 in maize. Mol. Breed. 28, 511-526.
550 Yi, N.,Xu, S., 2008. Bayesian lasso for quantitative trait loci mapping. Genetics 179, 1045-1055.

Yu, H., Xie, W., Wang, J., Xing, Y., Xu, C., Li, X., Xiao, J.,Zhang, Q., 2011. Gains in qtl detection using an ultra-high density snp map based on population sequencing relative to traditional rflp/ssr markers. PLoS One 6, e17595.

Zhou, Y., Zhang, Z., Bao, Z., Li, H., Lyu, Y., Zan, Y., Wu, Y., Cheng, L., Fang, Y., Wu, K., *et al.*, 2022. Graph pangenome captures missing heritability and empowers tomato breeding. Nature 606, 527-534.

Figure Legends

Fig. 1. The schematic workflow of the MVBLUP algorithm, which entails the adaptive learning of optimal weights reflecting the significance of each view via DE algorithm. These learned weights are then employed to construct a multi-view kinship matrix for the MVBLUP model.

Fig. 2. Systematic test results of MVBLUP on the Tomato332 dataset. **A**: Heatmaps of the genetic relationships based on SNP, InDel, and SV. **B**: Prediction accuracy of LASSO, XGBoost, and BLUP using single-view data. **C**: Prediction accuracy of BLUP using single-view data and MVBLUP integrating two or three views data. **D**: Prediction accuracy of five methods using three views data. BLUP_W1: integrating multi-view data with uniform weights and utilizing BLUP for phenotypic prediction; BLUP_W2: integrating multi-view data with weights determined by the average accuracy of five-fold cross-validation on single-view data training sets and employing BLUP for phenotypic prediction.

Fig. 3. Prediction accuracy of MVBLUP and BLUP using single-view data on the Rice210 dataset, for traits including grain number per panicle (**A**), 1000 grain weight (**B**), yield per plant trait (**C**), and tiller number per plant (**D**). G, genomic data; E, gene expression data; M, metabolomic data.

Fig. 4. Prediction accuracy of MVBLUP and BLUP using single-view data on the Maize368 dataset, for traits including heading date (**A**), silking time (**B**), pollen shedding (**C**), cob diameter (**D**), ear diameter (**E**), ear length (**F**), and ear leaf width (**G**), ear leaf length (**H**). G, genomic data; E, gene expression data; M, metabolomic data.

Fig. 5. Prediction accuracy of MVBLUP and BLUP using single-view data on the Maize282 dataset, for traits including days to anthesis (**A**), plant height (**B**), ear height (**C**), node number below ears (**D**), leaf width (**E**), and weight of 20 kernels (**F**). G, genomic data; E1–E7, gene

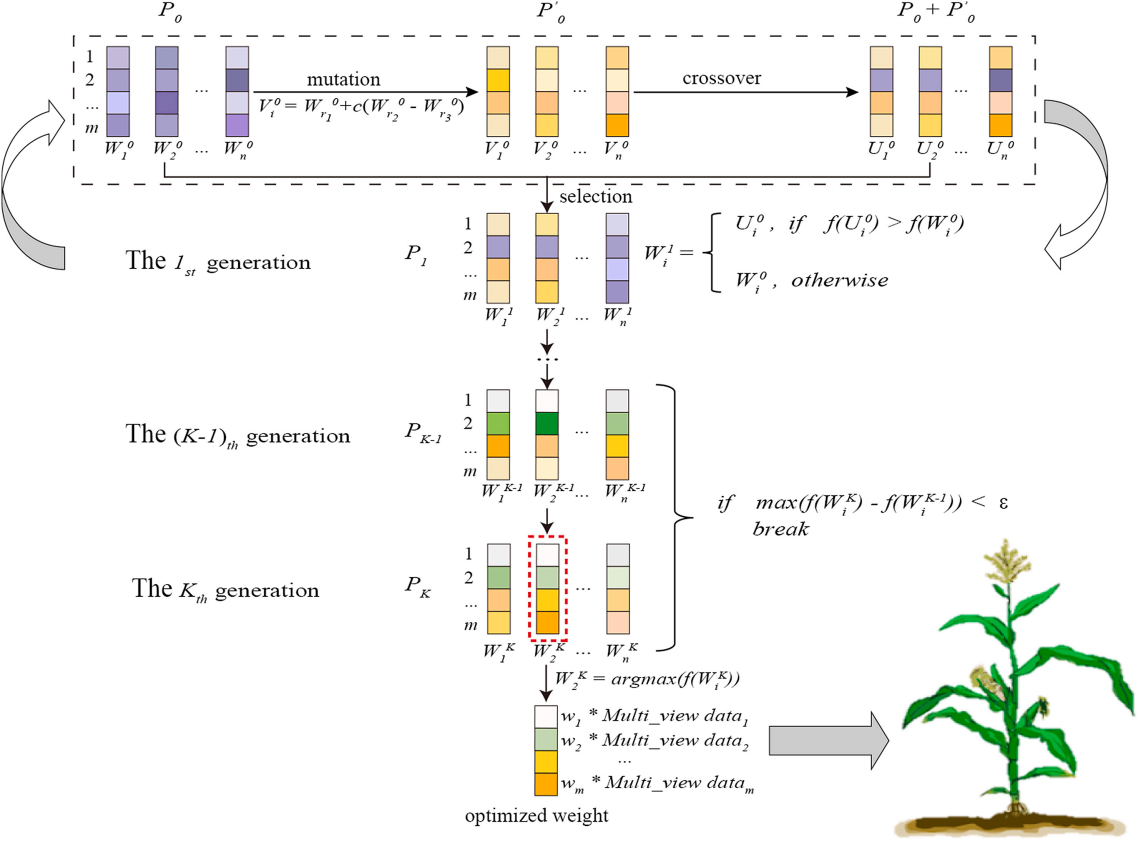
585 expression data from seven tissues, respectively, including germinating root, germinating shoot,
586 third leaf from the base, third leaf from the top, adult leaf collected during the day, adult leaf
587 collected at night, and mature kernel.

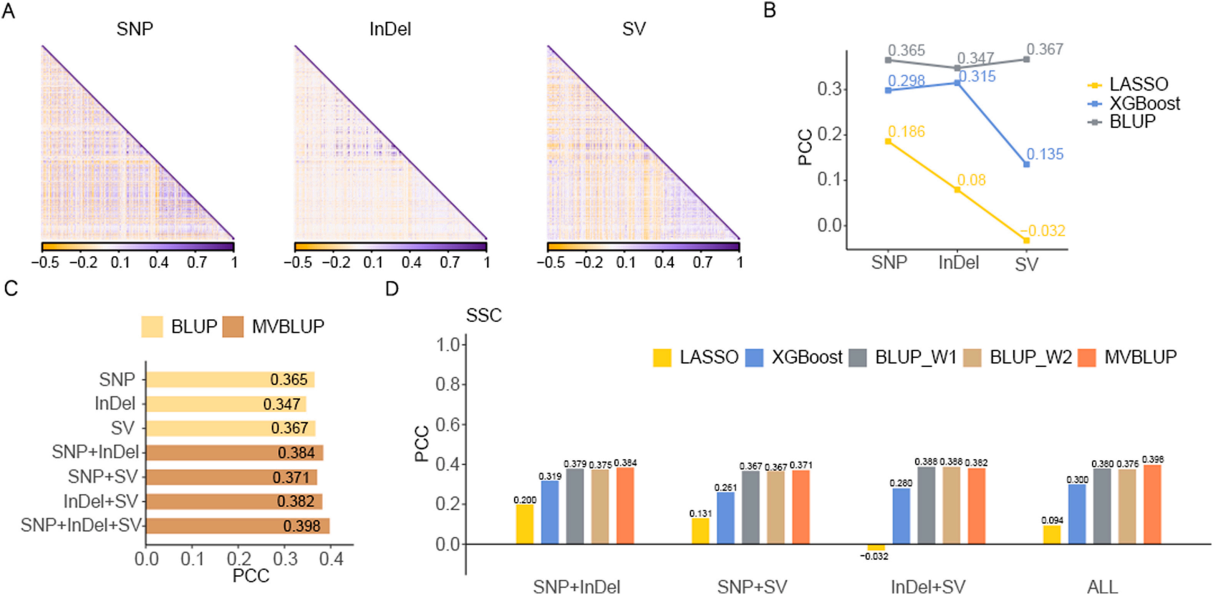
588

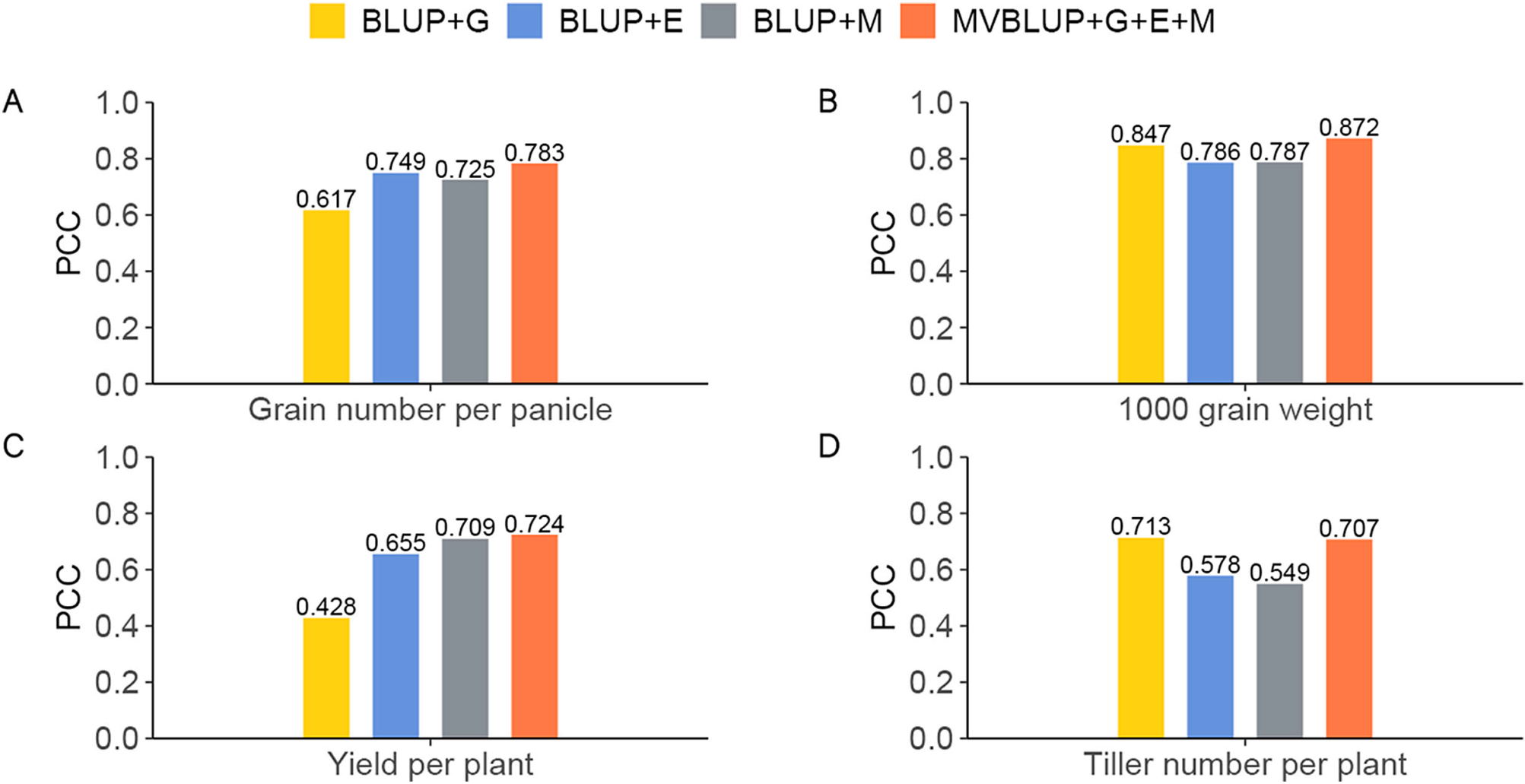
589 **Fig. 6.** Changes in prediction accuracy and weights during the learning process of the MVBLUP
590 algorithm. **A**: The training process of integrating eight single-view data sets G, E1, E2, E3, E4, E5,
591 E6, and E7 from Maize282 using the MVBLUP method, exemplified by the maize flowering days
592 phenotype. **B**: The changes in optimal weights of the eight views data during the training process.
593 G, genomic data; E1–E7, gene expression data from seven tissues respectively, including
594 germinating root, germinating shoot, third leaf from the base, third leaf from the top, adult leaf
595 collected during the day, adult leaf collected at night, and mature kernel.

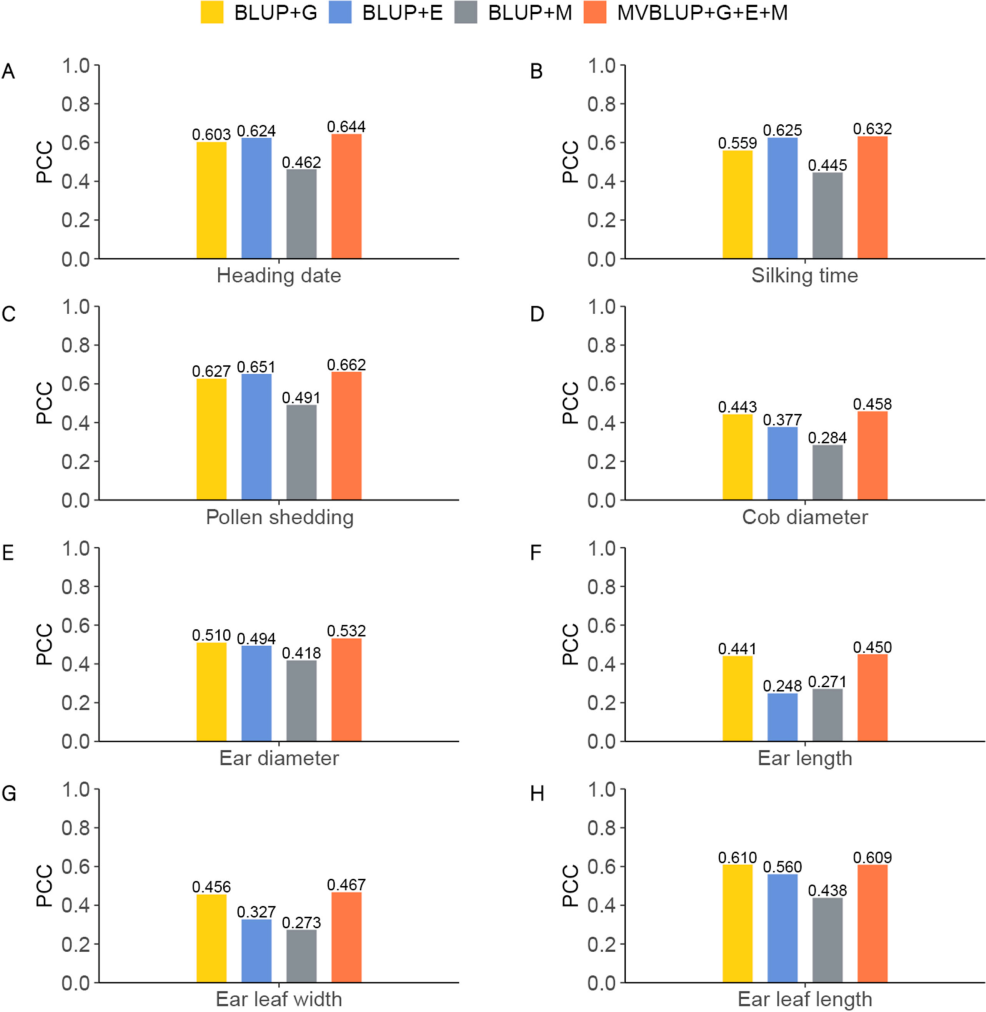
596

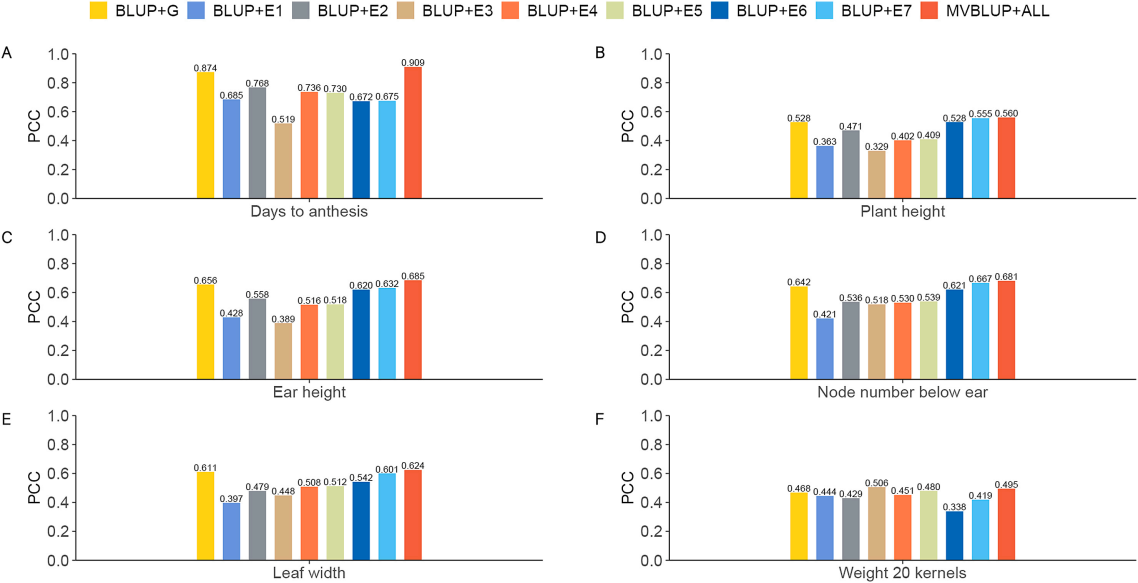
597

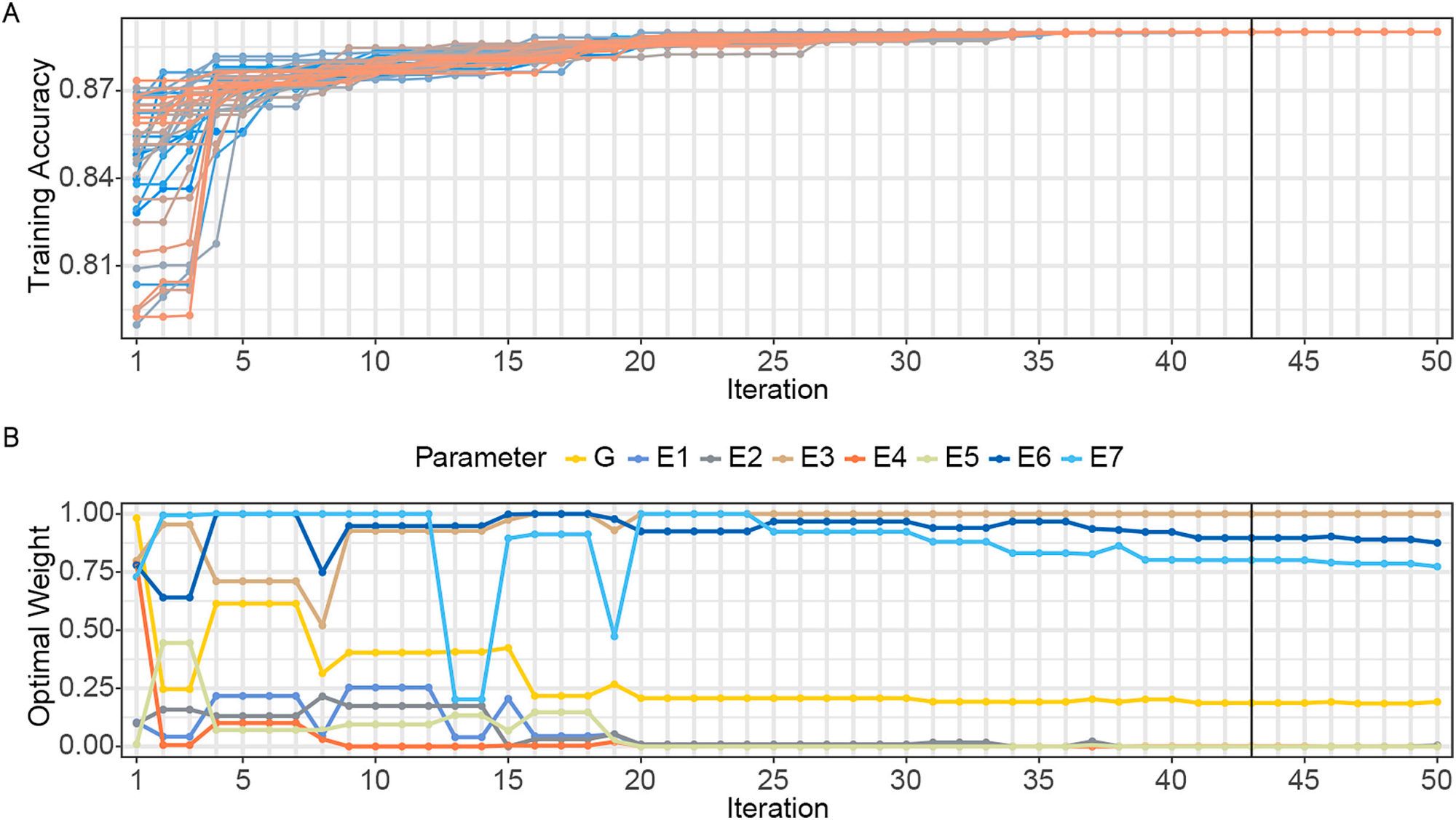

P_0
P'_0
$P_0 + P'_0$
mutation
$V_i^0 = W_{r_1}^0 + c(W_{r_2}^0 - W_{r_3}^0)$
crossover
selection
The 1_{st} generation
P_1
$W_i^1 = \begin{cases} U_i^0, & if \quad f(U_i^0) > f(W_i^0) \\ W_i^0, & otherwise \end{cases}$
The $(K\text{-}1)_{th}$ generation
P_{K-1}
The K_{th} generation
P_K
if $max(f(W_i^K) - f(W_i^{K-1})) < \varepsilon$
break
$W_2^K = argmax(f(W_i^K))$
w_1 * Multi_view data$_1$
w_2 * Multi_view data$_2$
w_m * Multi_view data$_m$
optimized weight


A
SNP
InDel
SV
−0.5 −0.2 0.1 0.4 0.7 1
B
PCC
0.365
0.347
0.367
0.298
0.315
0.135
0.186
0.08
−0.032
SNP
InDel
SV
LASSO
XGBoost
BLUP
C
BLUP
MVBLUP
SNP
InDel
SV
SNP+InDel
SNP+SV
InDel+SV
SNP+InDel+SV
0.365
0.347
0.367
0.384
0.371
0.382
0.398
PCC
D
SSC
LASSO
XGBoost
BLUP_W1
BLUP_W2
MVBLUP
PCC
0.200
0.319
0.379
0.375
0.384
0.131
0.261
0.367
0.367
0.371
−0.032
0.280
0.388
0.388
0.382
0.094
0.300
0.380
0.376
0.398
SNP+InDel
SNP+SV
InDel+SV
ALL


BLUP+G
BLUP+E
BLUP+M
MVBLUP+G+E+M
A
PCC
1.0
0.8
0.6
0.4
0.2
0.0
0.617
0.749
0.725
0.783
Grain number per panicle
B
PCC
0.847
0.786
0.787
0.872
1000 grain weight
C
PCC
0.428
0.655
0.709
0.724
Yield per plant
D
PCC
0.713
0.578
0.549
0.707
Tiller number per plant


BLUP+G
BLUP+E
BLUP+M
MVBLUP+G+E+M
A
PCC
1.0
0.8
0.6
0.4
0.2
0.0
0.603
0.624
0.462
0.644
Heading date
B
0.559
0.625
0.445
0.632
Silking time
C
0.627
0.651
0.491
0.662
Pollen shedding
D
0.443
0.377
0.284
0.458
Cob diameter
E
0.510
0.494
0.418
0.532
Ear diameter
F
0.441
0.248
0.271
0.450
Ear length
G
0.456
0.327
0.273
0.467
Ear leaf width
H
0.610
0.560
0.438
0.609
Ear leaf length


BLUP+G
BLUP+E1
BLUP+E2
BLUP+E3
BLUP+E4
BLUP+E5
BLUP+E6
BLUP+E7
MVBLUP+ALL
A
PCC
1.0
0.8
0.6
0.4
0.2
0.0
0.874
0.685
0.768
0.519
0.736
0.730
0.672
0.675
0.909
Days to anthesis
B
PCC
0.528
0.363
0.471
0.329
0.402
0.409
0.528
0.555
0.560
Plant height
C
PCC
0.656
0.428
0.558
0.389
0.516
0.518
0.620
0.632
0.685
Ear height
D
PCC
0.642
0.421
0.536
0.518
0.530
0.539
0.621
0.667
0.681
Node number below ear
E
PCC
0.611
0.397
0.479
0.448
0.508
0.512
0.542
0.601
0.624
Leaf width
F
PCC
0.468
0.444
0.429
0.506
0.451
0.480
0.338
0.419
0.495
Weight 20 kernels


A
Training Accuracy
0.87
0.84
0.81
1
5
10
15
20
25
30
35
40
45
50
Iteration
B
Parameter
G
E1
E2
E3
E4
E5
E6
E7
Optimal Weight
1.00
0.75
0.50
0.25
0.00
1
5
10
15
20
25
30
35
40
45
50
Iteration