OXFORD

Briefings in Bioinformatics, 2025, **26(1)**, bbaf058
https://doi.org/10.1093/bib/bbaf058
Problem Solving Protocol

mmGEBLUP: an advanced genomic prediction scheme for genetic improvement of complex traits in crops through integrative analysis of major genes, polygenes, and genotype–environment interactions

Qi-Xin Zhang[1,2], Tianneng Zhu[1], Feng Lin[1], Dunhuang Fang[2], Xuejun Chen[2], Xiangyang Lou[3], Zhijun Tong[2,*], Bingguang Xiao[2,*], Hai-Ming Xu[1,*]

[1]Institute of Crop Science and Institute of Bioinformatics, Zhejiang University, 866 Yuhangtang Road, Xihu District, Hangzhou, Zhejiang 310058, China
[2]Key Laboratory of Tobacco Biotechnological Breeding, National Tobacco Genetic Engineering Research Center, Yunnan Academy of Tobacco Agricultural Sciences, 33 Yuantong Road, Wuhua Distrct, Kunming, Yunnan 650021, China
[3]Department of Biostatistics, University of Florida, Gainesville, FL 32611, United States

*Corresponding authors. Zhijun Tong, Key Laboratory of Tobacco Biotechnological Breeding, National Tobacco Genetic Engineering Research Center, Yunnan Academy of Tobacco Agricultural Sciences, Kunming, Yunnan 650021, China. E-mail: tzj861@163.com; Bingguang Xiao, Key Laboratory of Tobacco Biotechnological Breeding, National Tobacco Genetic Engineering Research Center, Yunnan Academy of Tobacco Agricultural Sciences, Kunming, Yunnan 650021, China. E-mail: xiaobgkm@sina.com; Hai-Ming Xu, Institute of Crop Science and Institute of Bioinformatics, Zhejiang University, Hangzhou, Zhejiang 310058, China. E-mail: hmxu@zju.edu.cn

Abstract

Current genomic prediction (GP) models often fall short of fully capturing the genetic architecture of complex traits and providing practical breeding guidance, particularly under varying environments. Here, we propose the mmGEBLUP, an advanced GP scheme designed to tackle the current limitations in fully exploiting the genetic architecture of complex traits and to predict individual breeding value (BV) with multi-environment trial data. Our approach considers four genetic structural indicators to capture the genetic architectures stepwise across four models: the Genomic Best Linear Unbiased Prediction (GBLUP) model considers only main polygenic effects; the GEBLUP model includes both main and genotype-by-environment (GE) interaction polygenic effects; and the mmGBLUP and mmGEBLUP models further incorporate main and GE interaction effects of major genes. Through systematic simulations and applications to nine traits, three in rice and six in tobacco, we show stepwise increases in prediction accuracy from GBLUP to mmGEBLUP, providing evidence on the scale of heritability and polygenicity of traits. In practical terms, we predict four components of BV: major additive, minor additive, major interaction, and minor interaction. Interestingly, we discover that for traits like natural leaf number in tobacco, the major additive BVs for the top 20 individuals are substantially equal; it is the minor additive BV that causes the difference in the total BV. The relative size of major/minor additive BVs suggests performing either marker-assisted selection or genomic selection or both. Overall, mmGEBLUP is an advanced prediction scheme that enhances the understanding of genetic architectures and facilitate the genetic improvement of complex traits in crops under diverse environments.

Keywords: genomic prediction; complex traits; genetic architecture; GE interaction; breeding value

Introduction

Over the past two decades, genomic selection (GS) or genomic prediction (GP) has significantly enhanced genetic gains and accelerated crop breeding. GP leverages phenotypic and genomic data from a training population to estimate breeding values (BVs) for individuals in a validation or testing population that have been genotyped but not phenotyped. With its efficiency, accuracy, and rationality, GP continues to advance and improve in the era of Breeding 4.0, with considerably high accuracy [1, 2]. However, as breeding elite varieties under continuous selection for complex traits approaches saturation, more advanced methods are essential for further improvement.

The performance of GP models depends on how well they capture the genetic architecture of traits. While plenty of GP methods exist, the Genomic Best Linear Unbiased Prediction (GBLUP) model is a commonly used reference due to its simplicity, versatility in incorporating diverse information, and robustness across traits [3]. Despite its strengths, GBLUP assumes equal genetic variance for all single-nucleotide polymorphisms (SNPs) and considers only additive effects. Trait genetic architecture varies not only by heritability but also by polygenicity, which reflects the proportion of genetic variance attributed to large-effect genes (major genes) versus small-effect genes (polygenes or minor genes). In crops like rice, genome-wide association studies (GWASs) have

Received: October 11, 2024. **Revised:** December 24, 2024. **Accepted:** January 28, 2025

identified quantitative trait loci (QTLs) with large effects on traits such as grain yield, flowering time, plant height [4], and aluminum tolerance [5], while in crops like maize, studies have found most agronomic traits to be controlled by numerous small-effect genes, such as leaf structure [6] and flowering time [7]. Similarly, in tobacco, linkage analysis in biparental and multiparent advanced generation intercross (MAGIC) populations has revealed that traits like leaf length are influenced by both major and minor QTLs [8, 9].

To address the complexity of traits with varying polygenicity, GBLUP has been modified to incorporate peak marker–trait associations. These modifications include weighting causal mutations in the genetic relationship matrix (GRM) [10], treating associated markers as fixed effects [11–15], and adding random effects for GWAS signals [16]. Such augmented GP models improve predictions for traits mostly influenced by large-effect genes, such as flowering time or disease resistance [12, 13]. The problem is how to identify variants with large causal effects. Previous studies have evaluated GWAS models for GP using top variants based on *P*-values [17, 18] or fitness evaluation [19], but few have tested all combinations of significant signals for optimal fits [12]. To remind, increasing prediction accuracy (PA) is not the ultimate goal; rather, the predicted individual BV is what truly matters in breeding programs.

Genotype-by-environment (GE) interaction is a fundamental aspect of biology and plays a crucial role in shaping the genetic architecture of complex traits [20]. Plant breeding usually includes extensive multilocation and multiyear field trial data, which served as a great source for analyzing genotypic variation across environments. GP is popular in predicting genotypic values in annual crops, where both genetic and nongenetic effects determine the final commercial value of the lines [21–23]. Multi-environment GP models have been successfully implemented in various crops such as rice [24], maize [25], and wheat [26]. However, while core variants have been incorporated into GP, no existing method fully integrates major and minor additive and interaction effects for multi-environment trial data. This highlights the need for new prediction methods tailored to both genetic architecture and GE interaction structures.

To address these limitations, we propose an advanced GP scheme that exploits the genetic architecture of complex traits and provides practical breeding guidance, especially under diverse environmental conditions. The scheme, outlined in Fig. 1, includes three key highlights: (i) we use four genetic structural indicators to describe the genetic architecture and GE structure of traits. The first two indicators—general and interaction heritability—are well known in quantitative genetics. The other two indicators—general and interaction polygenicity—are introduced to crop breeding for the first time in this study. (ii) We include four GP models—GBLUP, GEBLUP, mmGBLUP, and mmGEBLUP—as part of the scheme. Each model incrementally incorporates one additional genetic structural indicator. Through simulations and applications, we show how stepwise increases in PA reveal heritability and polygenicity scales for different traits. (iii) Traditional BV is defined as the additive genetic values. In our study, mmGEBLUP predicts four BV components: major additive, minor additive, major interaction, and minor interaction. The relative contributions of major and minor BVs inform whether to use marker-assisted selection (MAS) or GS or both. Additionally, interaction term predictions guide breeding for environment-specific or broadly adaptive varieties.

Materials and methods

Statistical models

In essence, mmGEBLUP is established on the empirical knowledge of the genetic architecture of traits, including GE interactions. The main idea of introducing four GP models, GBLUP, GEBLUP, mmGBLUP, and mmGEBLUP, as a whole system is to step-wisely break down the genetic architecture of the traits. Therefore, the features and the assumptions of the model are also step-wisely brought into the model (Fig. 1 and Table S1). We will introduce model assumptions and notations step by step, starting from the basic GBLUP model.

Genomic Best Linear Unbiased Prediction model

Let's suppose we have observations across v environments for g lines. Based on the GBLUP model, the phenotypic value of the i-th line in the h-th environment (y_{ih}) can be expressed by the following mixed linear model:

$$y_{ih} = \mu + a_i + e_h + \varepsilon_{ih}$$

where μ is the population mean; a_i is the additive genetic effect of i-th line, random effect, $a_i \sim N(0, \sigma_A^2)$; e_h is the main effect of the h-th environment, random effect, $e_h \sim N(0, \sigma_E^2)$; and ε_{ih} is the residual effect, random effect, $\varepsilon_{ih} \sim N(0, \sigma_\varepsilon^2)$. The model can also be written in matrix form,

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Z}_A\mathbf{u}_A + \mathbf{Z}_E\mathbf{u}_E + \boldsymbol{\varepsilon} \quad (1)$$

where $\mathbf{y}$ is an $n \times 1$ vector of phenotypic values; n is the total number of observations, where $n = g \times v$. Incomplete data with fewer observations than n can also be analyzed by modifying the design matrices; $\boldsymbol{\mu} = [\mu \ \ldots \ \mu]'$ is the vector of population mean; $\mathbf{u}_A = [a_1 \ \ldots \ a_g]' \sim MVN(0, \mathbf{K}_A\sigma_A^2)$ with the coefficient matrix $\mathbf{Z}_A = \mathbf{1}_v \otimes \mathbf{I}_g$; $\otimes$ denotes the Kronecker product of matrices; $\mathbf{1}_v$ denotes an $v \times 1$ vector with all the elements equal to 1, and $\mathbf{I}_g$ denotes an $g \times g$ identity matrix; $\mathbf{K}_A$ is the additive relationship matrix estimated by Endelman's method [27], $\mathbf{K}_A = \frac{\mathbf{MM}'}{2\sum_{j=1}^{m} p_j(1-p_j)}$, where p_j is the minor allele frequency of the j-th marker and m is the number of markers. $\mathbf{M} = \{x_{ij}\}$ is the genotype matrix, whose element is coded with −1, 0, and 1 for the null homozygous, heterozygote, and positive homozygote, respectively, and centered by subtracting $2p_j - 1$; $\mathbf{u}_E = [e_1 \ \ldots \ e_v]' \sim MVN(0, \mathbf{I}\sigma_E^2)$ with the coefficient matrix $\mathbf{Z}_E = \mathbf{I}_v \otimes \mathbf{1}_g$; $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of residual effects, $\boldsymbol{\varepsilon} \sim MVN(0, \mathbf{I}_n\sigma_\varepsilon^2)$. The GBLUP model only considers one genetic structural indicator, which is the general heritability.

Genotype–Environment Best Linear Unbiased Prediction model

To further incorporate GE interaction, we adopted the GEBLUP (Genotype–Environment Best Linear Unbiased Prediction) model as an extension from the basic GBLUP model. Based on the GEBLUP model, the phenotypic value y_{ih} can be expressed as,

$$y_{ih} = \mu + a_i + e_h + ae_{ih} + \varepsilon_{ih}$$

where ae_{ih} is the additive-by-environment interaction effect due to all markers of the i-th line and h-th environment, random effect, $ae_{ih} \sim N(0, \sigma_{AE}^2)$. Other parameters denoted in the model are the same as those in the GBLUP model. The model can also be written

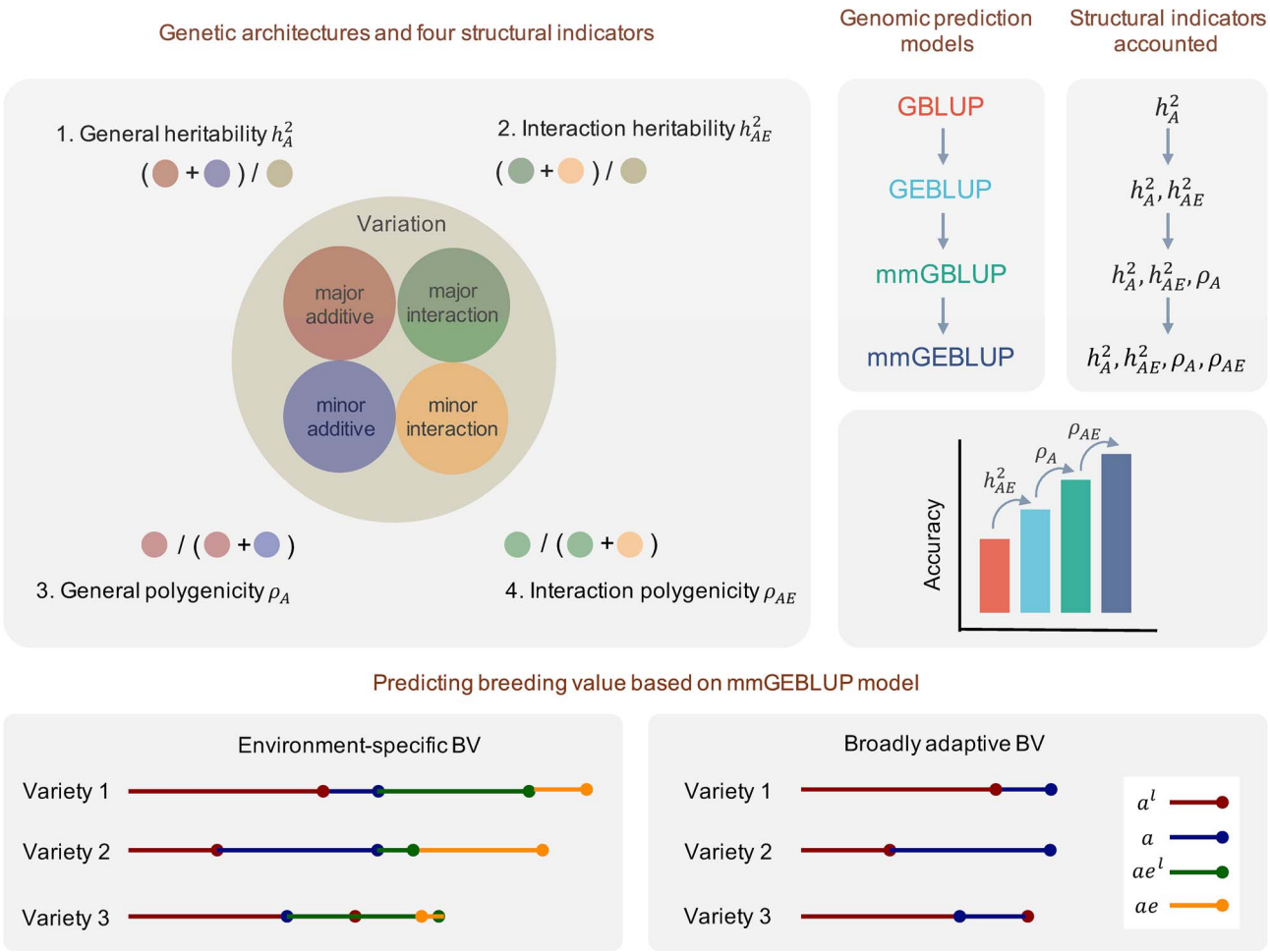

Figure 1. Description of the mmGEBLUP scheme. Genetic architectures and four structural indicators: the basic idea of the mmGEBLUP model lies in the phenotypic variance explained by different genetic resources and their relative sizes. We adopt four structural indicators to describe genetic architectures and GE interaction structures. h_A^2 is the general heritability in the narrow sense; h_{AE}^2 is the interaction heritability in the narrow sense; ρ_A is the general polygenicity that denotes the proportion of additive variance explained by markers with major additive effects; ρ_{AE} is the interaction polygenicity that denotes the proportion of additive-by-environment interaction variance explained by markers with major interaction effects; genomic prediction models and structural indicators accounted. We compare four GP models that have different assumptions on marker effects and model different genetic architectures, which are GBLUP, GEBLUP, mmGBLUP, and mmGEBLUP. The relative size of prediction accuracy between models is related to different genetic structural indicators; predicting BVs based on the mmGEBLUP model: environment-specific and broadly adaptive BVs can be predicted for each individual based on four components, which are major additive BV (a^l), minor additive BV (a), major interaction BV (ae^l), and minor interaction BV (ae).

in matrix form,

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Z}_A\mathbf{u}_A + \mathbf{Z}_E\mathbf{u}_E + \mathbf{Z}_{AE}\mathbf{u}_{AE} + \boldsymbol{\varepsilon} \quad (2)$$

where $\mathbf{u}_{AE} = [ae_{11} \cdots ae_{g1} \ldots ae_{1v} \ldots ae_{gv}]' \sim MVN\left(0, \mathbf{I}_v \otimes \mathbf{K}_A\sigma_{AE}^2\right)$ with the coefficient matrix $\mathbf{Z}_{AE} = \mathbf{I}_v \otimes \mathbf{I}_g$. The GBLUP model has been extended to GEBLUP, which implies both general and interaction heritability, yet it still overlooks the polygenic nature of these traits and their interactions with the environment.

Major–minor integrated Genomic Best Linear Unbiased Prediction model

By introducing different assumptions regarding the impact of additive marker effects, we then describe the major–minor integrated Genomic Best Linear Unbiased Prediction (mmGBLUP) model. In the mmGBLUP model, the additive effects for causal markers are assumed as fixed effects, while the additive effects for the remaining markers are assumed as random effects following the same distribution. Suppose there are t causal markers with large additive effects. Based on the mmGBLUP model, the phenotypic value can be expressed as,

$$y_{ih} = \mu + \sum_{j=1}^{t} x_{ij}q_j + a_i + e_h + ae_{ih} + \varepsilon_{ih}$$

$$= \mu + a_i^l + a_i + e_h + ae_{ih} + \varepsilon_{ih}$$

where $a_i^l = \sum_{j=1}^{t} x_{ij}q_j$ is the summation of large additive effects for i-th line, fixed effect; x_{ij} is the coding value (−1, 0, or 1) of the genotype, and the q_j is the additive effect; a_i is the additive genetic effect of the i-th line due to the remaining small-effect markers, random effect, $a_i \sim N\left(0, \sigma_A^2\right)$. Other parameters denoted in the model are the same as those in the GEBLUP model. The model can also be written in matrix form,

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{X}_{A^l}\mathbf{b}_{A^l} + \mathbf{Z}_A\mathbf{u}_A + \mathbf{Z}_E\mathbf{u}_E + \mathbf{Z}_{AE}\mathbf{u}_{AE} + \boldsymbol{\varepsilon} \quad (3)$$

where $\mathbf{b}_{A^l} = \begin{bmatrix} q_1 & \dots & q_t \end{bmatrix}'$ is an $t \times 1$ effect vector for large-effect markers with the coefficient matrix $\mathbf{X}_{A^l} = \mathbf{1}_v \otimes \mathbf{M}^l$; $\mathbf{M}^l = \begin{bmatrix} \mathbf{x}_1 & \dots & \mathbf{x}_g \end{bmatrix}'$ is a $g \times t$ genotype matrix consisting of $\mathbf{x}_i = \begin{bmatrix} x_{i1} & \dots & x_{it} \end{bmatrix}'$ $(i = 1, \dots, g)$, which is a $t \times 1$ genotype vector of the i-th line on t large-effect markers; $\mathbf{u}_A = \begin{bmatrix} a_1 & \dots & a_g \end{bmatrix}' \sim MVN(0, \mathbf{K}_A\sigma_A^2)$ with the coefficient matrix $\mathbf{Z}_A = \mathbf{1}_v \otimes \mathbf{I}_g$; $\mathbf{K}_A$ is the additive relationship matrix constructed with small-effect background. Other parameters are the same as the denoted in the GEBLUP model. Extending from GEBLUP model, the mmGBLUP model accounts for an extra genetic structural indicator, the general polygenicity.

Major–minor integrated Genotype-Environment Best Linear Unbiased Prediction model

By introducing different assumptions regarding the impact of large and small GE interaction effects, we propose the major–minor integrated Genotype–Environment Best Linear Unbiased Prediction (mmGEBLUP) model. This model introduces different assumptions about the contributions of large and small GE interaction effects. In mmGEBLUP, additive-by-environment interaction effects for different causal markers are modeled as random effects with distinct variances, while interaction effects for the remaining markers are treated as random effects following a common distribution. Suppose there are t causal markers with large additive and s causal markers with large additive-by-environment interaction effects. Based on the mmGEBLUP model, the phenotypic value can be expressed as,

$$\begin{aligned} y_{ih} &= \mu + \sum_{j=1}^{t} x_{ij}q_j + a_i + e_h + \sum_{j=1}^{s} x_{ij}(qe)_{jh} + ae_{ih} + \varepsilon_{ih} \\ &= \mu + a_i^l + a_i + e_h + ae_{ih}^l + ae_{ih} + \varepsilon_{ih} \end{aligned}$$

where μ, a_i^l, a_i, e_h, and ε_{ih} are denoted the same as those in mmGBLUP model; $ae_{ih}^l = \sum_{j=1}^{s} x_{ij}(qe)_{jh}$ is the summation of large additive-by-environment marker effects for i-th line in the h-th environment; $(qe)_{jh}$ is the additive-by-environment interaction effect of the j-th large-effect marker in the h-th environment, random effect, $(qe)_{jh} \sim N\left(0, \sigma_{A_jE}^2\right)$; ae_{ih} is the additive-by-environment interaction effect due to all small-effect markers for the i-th line in the h-th environment, random effect, $ae_{ih} \sim N(0, \sigma_{AE}^2)$. The model can also be written in matrix form,

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{X}_{A^l}\mathbf{b}_{A^l} + \mathbf{Z}_A\mathbf{u}_A + \mathbf{Z}_E\mathbf{u}_E + \mathbf{Z}_{A^lE}\mathbf{u}_{A^lE} + \mathbf{Z}_{AE}\mathbf{u}_{AE} + \boldsymbol{\varepsilon} \quad (4)$$

where $\mathbf{u}_{A^lE} = \begin{bmatrix} (qe)_{11} & \dots & (qe)_{s1} & \dots & (qe)_{1v} & \dots & (qe)_{sv} \end{bmatrix}' \sim MVN\left(0, \mathbf{I}_v \otimes \mathrm{diag}\left(\sigma_{A_1E}^2 \quad \dots \quad \sigma_{A_sE}^2\right)\right)$ with the coefficient matrix $\mathbf{Z}_{A^lE} = \mathbf{I}_v \otimes \mathbf{X}_{A^l}$; $\mathbf{u}_{AE} = \begin{bmatrix} ae_{11} & \dots & ae_{g1} & \dots & ae_{1v} & \dots & ae_{gv} \end{bmatrix}' \sim MVN(0, \mathbf{I}_v \otimes \mathbf{K}_A\sigma_{AE}^2)$ with the coefficient matrix $\mathbf{Z}_{AE} = \mathbf{I}_v \otimes \mathbf{I}_g$; $\mathbf{K}_A$ is constructed with the remaining markers with small interaction effects. Other parameters are denoted the same as those in the mmGBLUP model. So far, we have described the mmGEBLUP model along with other three GP models and how these model assumptions account for different genetic architecture.

The prediction of effects and breeding values

Models are fitted using the *sommer* R package [28]. All four models are built under the framework of mixed linear model. The typical mixed models developed by Searle and Henderson can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon} \quad (5)$$

where $\boldsymbol{\beta}$ and $\mathbf{u}$ are vectors of the fixed and random effects, respectively. $\mathbf{X}$ and $\mathbf{Z}$ are the incidence matrices. For example, in the mmGEBLUP model, the fixed and random effects are $\boldsymbol{\beta} = \begin{bmatrix} \mu \\ \mathbf{b}_{A^l} \end{bmatrix}$ and $\mathbf{u} = \begin{bmatrix} \mathbf{u}_A \\ \mathbf{u}_E \\ \mathbf{u}_{A^lE} \\ \mathbf{u}_{AE} \end{bmatrix}$ and the corresponding incidence matrices are $\mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{X}_{A^l} \end{bmatrix}$ and $\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_A & \mathbf{Z}_E & \mathbf{Z}_{A^lE} & \mathbf{Z}_{AE} \end{bmatrix}$. The variance–covariance structure of $\mathbf{y}$ is $\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$, where $\mathbf{G}$ is the variance–covariance matrix of random effect vector $\mathbf{u}$ and $\mathbf{R}$ is the variance–covariance matrix of residual effect vector $\boldsymbol{\varepsilon}$. Based on the direct-inversion method, the estimators of fixed and random effect are $\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\mu} & \hat{\mathbf{b}}_{A^l} \end{bmatrix}' = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ and $\hat{\mathbf{u}} = \begin{bmatrix} \hat{\mathbf{u}}_A' & \hat{\mathbf{u}}_E' & \hat{\mathbf{u}}_{A^lE}' & \hat{\mathbf{u}}_{AE}' \end{bmatrix}' = \mathbf{GZ}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. The estimation of both $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ requires the knowledge of variance–covariance parameters. To obtain estimates of these parameters, restricted maximum likelihood based on the Newton iterative algorithm is used to estimate the variance components via the *sommer* R package. We are able to predict the total genetic value of i-th individual within h-th environment through four BV components, based on their origins, referred to as major additive BV (a^l), minor additive BV (a), major interaction BV (ae^l), and minor interaction BV (ae), respectively. The prediction of these four BV components is based on the estimation of fixed and random effects,

$$\begin{bmatrix} \hat{a}_1^l \\ \vdots \\ \hat{a}_g^l \end{bmatrix} = \mathbf{X}_{A^l}\hat{\mathbf{b}}_{A^l}, \begin{bmatrix} \hat{a}_1 \\ \vdots \\ \hat{a}_g \end{bmatrix} = \hat{\mathbf{u}}_A, \begin{bmatrix} \hat{ae}_{11}^l \\ \vdots \\ \hat{ae}_{gv}^l \end{bmatrix} = \mathbf{Z}_{A^lE}\hat{\mathbf{u}}_{A^lE}, \begin{bmatrix} \hat{ae}_{11} \\ \vdots \\ \hat{ae}_{gv} \end{bmatrix} = \hat{\mathbf{u}}_{AE}$$

Model evaluation

We employed the cross-validation (CV) scheme to evaluate the performance of models. For example, in a k-fold cross-validation experiment, one of the k folds served as the validation fold, and the other k-1 folds served as training folds. This process was repeated k times so that each fold served once as the validation fold. Given the multi-environment experiment design, we defined that the training population includes all folds in training environments and training folds in validation environments, while the validation population consisted only of the validation fold in the validation environment (Fig. 2).

The means and standard errors of PA, calculated from Pearson's correlation between the observed and the predicted phenotype, will be reported and pair-wise Student's t-test is employed to test whether the performance of one method is better than that of the other. Let x_{1i} denote the PA of one method and x_{0i} denote the PA of another method. The difference in PA between the two methods under i-th CV is $d_i = x_{1i} - x_{0i}$, $\bar{d}$ is the mean difference of d_i. The t-statistic is

$$T^* = \frac{\bar{d}}{SE\left(\bar{d}\right)} \quad (6)$$

Under the null hypothesis H_0: $\mu_1 \leq \mu_0$ and alternative hypothesis H_1: $\mu_1 > \mu_0$, this statistic follows a Student t-distribution with $n - 1$ degree of freedom. If it is a k-fold cross-validation, with c repeats, then $n = kc$.

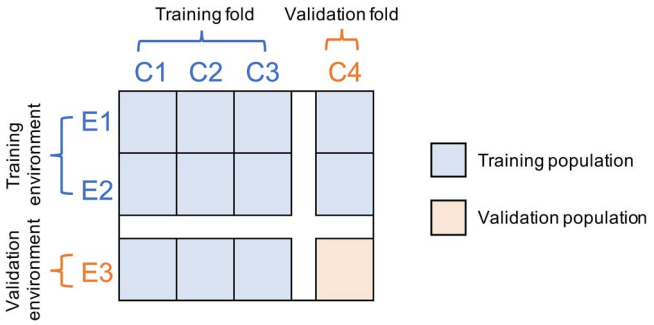

Figure 2. Description of the training and validation populations under multi-environment trials. There are three environments (E1, E2, and E3) and four CV folds (C1, C2, C3, and C4) in the example. The training population includes all folds in training environments and only training folds in the validation environment, while the validation population includes only the validation fold in the validation environment.

Genetic mapping approaches

To apply the mmGEBLUP method more effectively, it is crucial to identify markers that are closely associated with causal genes. Previous studies have investigated different GWAS models to be implemented in GP. Most of them chose to incorporate top signals sorted by *P*-values [17, 18]. Few have adopted further tests on all combinations of significant signals for optimal fit [12]. Here, we adopted a similar strategy to detect causal markers with additive or additive-by-environment effects in complex traits. We employed the mixed-linear-model-based QTXNetwork 2.0 software to select significant signals and remain significant ones in the stepwise model selection [29]. To control the experiment-wise type I error rate, a critical *F*-value based on the Henderson III method was determined by the permutation test with 1000 repeats for each tested locus at the significance level of .05. To remove spurious signals, the stepwise selection was performed on all significant signals from the *F*-statistic profile. All signals with either additive or additive-by-environment effects remained after stepwise selection will be considered as large-effect markers in mmGBLUP and mmGEBLUP models.

Simulation

To properly quantify the genetic architecture of traits, we proposed a systematic procedure of phenotype simulation according to four genetic instructors, which are general and interaction heritability and polygenicity. The definitions of these four instructors and a transformation from the four instructors to the generation of marker effects are described in the following context.

Here, we consider additive genetic variance, but the model can further include dominance and epistasis genetic variances conceptually. When there exists GE interaction, phenotypic variance (σ_P^2) is partitioned as $\sigma_P^2 = \sigma_A^2 + \sigma_E^2 + \sigma_{AE}^2 + \sigma_\varepsilon^2$. Here, σ_A^2 is the variance of additive genetic effects, σ_E^2 is the environmental variance, σ_{AE}^2 is the additive-by-environment interaction variance, and σ_ε^2 is the residual variance. Heritability including general heritability (h_G^2, here h_A^2) and interaction heritability (h_{GE}^2, here h_{AE}^2) in the narrow sense can be expressed as,

$$h_A^2 = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_{AE}^2 + \sigma_\varepsilon^2}$$

$$h_{AE}^2 = \frac{\sigma_{AE}^2}{\sigma_A^2 + \sigma_{AE}^2 + \sigma_\varepsilon^2} \quad (7)$$

Building on the work of Yang and Zhou [30], we adopt the concept of PGE (ρ), which is the proportion of genetic variance explained by the large-effect term to quantify the polygenic hypothesis fittings or the 'polygenicity' for a surveyed trait. We now generalize the concept of polygenicity to GE interaction. Let's assume we have m markers, with t of them being large-effect markers with the same variance for additive effects and s of them being large-effect markers with the same variance for additive-by-environment effects, respectively. The general polygenicity (ρ_A) and interaction polygenicity (ρ_{AE}) can be expressed as,

$$\rho_A = \frac{\sum_j^t (2p_j(1-p_j))\sigma_{a_l}^2}{\sigma_A^2} = 1 - \frac{\sum_j^{m-t} (2p_j(1-p_j))\sigma_{a_s}^2}{\sigma_A^2}$$

$$\rho_{AE} = \frac{\sum_j^s (2p_j(1-p_j))\sigma_{ae_l}^2}{\sigma_{AE}^2} = 1 - \frac{\sum_j^{m-s} (2p_j(1-p_j))\sigma_{ae_s}^2}{\sigma_{AE}^2} \quad (8)$$

where p_j is the allele frequency of j-th marker, $\sigma_{a_l}^2$ and $\sigma_{a_s}^2$ are the additive variances of marker effects whose contribution is relatively large or small; $\sigma_{ae_l}^2$ and $\sigma_{ae_s}^2$ are the interaction variances of marker effects whose contribution is relatively large or small. We simulate both the effects of major and minor markers by random sampling from a normal distribution. If we assume $\sigma_A^2 + \sigma_{AE}^2 + \sigma_\varepsilon^2 = 1$, based on Eq. (7) and Eq. (8), the variance of marker effects in terms of heritability (h_A^2 and h_{AE}^2) and polygenicity (ρ_A and ρ_{AE}) are

$$\sigma_{a_l}^2 = \frac{\rho_A h_A^2}{\sum_j^t (2p_j(1-p_j))} textrmand\ \sigma_{a_s}^2 = \frac{(1-\rho_A)h_A^2}{\sum_j^{m-t}(2p_j(1-p_j))}$$

$$\sigma_{ae_l}^2 = \frac{\rho_{AE} h_{AE}^2}{\sum_j^s (2p_j(1-p_j))} textrmand\ \sigma_{ae_s}^2 = \frac{(1-\rho_{AE})h_{AE}^2}{\sum_j^{m-s}(2p_j(1-p_j))} \quad (9)$$

Consequently, marker effects are simulated based on two parts, the main effect (q_j) and the interaction effect (qe_{jh}). For example, the main effect of j-th large-effect marker is randomly sampled from $q_j \sim N(0, \sigma_{a_l}^2)$, its interaction effect in the h-th environment is randomly sampled from $qe_{jh} \sim N(0, \sigma_{ae_l}^2)$, and its overall effect is $q_j + qe_{jh}$. Similarly, for small-effect markers, $\sigma_{a_l}^2$ is replaced by $\sigma_{a_s}^2$ and $\sigma_{ae_l}^2$ is replaced by $\sigma_{ae_s}^2$. We visualized the simulation for the main effect, interaction effect, and overall effect in Fig. S1, using three environments as an example.

In this study, we generated genotypes for 1000 individuals with 2000 independent markers and a total of 3000 observations across three environments. We randomly selected minor allele frequencies of markers from a uniform distribution of $U(0.05, 0.5)$. For more complicated simulation scenarios involving up to 100 000 markers and different LD conditions, please refer to SNote 1. For the phenotypic values, we designated seven markers as major causal variants with large additive or additive-by-environment effects or both, while the remaining markers were assumed as minor causal variants with small additive and additive-by-environment effects. To examine model performance under different genetic architectures and GE interaction structures, we considered a total of seven scenarios with different combinations of h_A^2, h_{AE}^2, ρ_A, and ρ_{AE} (Table S2). In each scenario, we conducted 4-fold cross-validation in 10 repeats and evaluated the performance of GBLUP, GEBLUP, mmGBLUP, and mmGEBLUP models.

Data description

We choose two breeding populations as examples for testing the effectiveness of employing mmGEBLUP in real data. To note that, these two populations represent different types of breeding; one is representative for diversity breeding panels, and the other is representative for a designed biparental population. Both populations are frequently used in multi-environment field trials.

Rice population

A rice diversity panel consisting of 369 elite inbred lines from the International Rice Research Institute (IRRI) irrigated rice breeding program was analyzed. 384-plex genotyping-by-sequencing was used to discover and call SNPs. SNPs with call rates <90% were removed along with monomorphic markers to obtain a filtered SNP dataset containing 73 147 SNPs. We further removed SNPs with low-frequency (MAF < 0.05) and high linkage disequilibrium (LD, $R^2 > 0.9$ in a 50-variant window and a 5-variant count to shift the window). A total of 8756 SNPs remained and were used in the following GP analysis. Individuals with missing data ≥60% and outlier individuals were removed according to principal components analysis as performed in Spindel *et al.* [31], and 332 lines remained for further analysis. Phenotypic data were obtained from a replicated yield trial conducted in Los Baños, Philippines, from 2009 to 2012 in two seasons per year: dry season (DS) and wet season (WS). These season-year combinations were considered as eight environments. Different training and validation environments assigned in different experiment designs are given in Table S3. Three phenotypes, namely, grain yield (YLD, kg/ha), plant height (PH, cm), and flowering time (FLW, days to 50% flowering) were analyzed. As there are differing degrees of family relatedness existed, we adopted the same familial background control before randomly assigning individuals to each CV fold [12, 31].

Tobacco population

A breeding population of 268 recombinant inbred lines (RILs) derived from K326 and Y3, two elite flue-cured tobacco parents, has been constructed to investigate the genetic basis of agronomic traits and hazardous smoke traits in tobacco. We considered 7107 bin markers mapped on 24 linkage groups based on an integrated SNP-Indel-SSR genetic map [9]. Six agronomic traits including natural plant height (nPH, cm), natural leaf number (nLN), stem girth (SG, cm), internode length (IL, cm), length of the largest leaf (LL, cm), and width of the largest leaf (LW, cm) were measured in Shilin from 2020 to 2022, denoted as 2020SL, 2021SL, and 2022SL. We considered 2022SL as the validation environment and 2020SL and 2021SL as the training environments. Other statistic details of the traits from two population are listed in Table S4.

Results

Simulated data

The bar plots depict the PA using GBLUP, GEBLUP, mmGBLUP, and mmGEBLUP models with different simulated traits under various genetic architectures and GE interaction structures (Fig. 3A–C). Different traits were simulated through diverse settings of general and interaction heritability and polygenicity (Table S2). PA was influenced by the genetic architectures of traits as well as the choice of GP methods. In general, the upper limit of PA was determined by trait heritability [32], which implied that higher heritability led to higher accuracy. For example, in Fig. 3A, PA from all four methods increased as general heritability increased. However, other genetic structural indicators, such as interaction heritability, considered in GP models further affected the upper limits. Since no GE interactions were considered in the GBLUP model, its PA remained relatively constant even as total heritability increased. For example, when $h^2_A = 0.3$ and h^2_{AE} increased from 0.2 to 0.3 and 0.6, the average PA of GBLUP is consistently 0.407, 0.405, and 0.398 (Fig. 3B). Unlike the GBLUP model, the upper limits of GEBLUP, mmGBLUP, and mmGEBLUP models that considered GE interaction depended on both general heritability and interaction heritability. For instance, the average accuracies for the GEBLUP model increased from 0.434 to 0.464 and 0.605 when h^2_{AE} increased from 0.2 to 0.3 and 0.6 (Fig. 3B).

As an advanced GP model, the mmGEBLUP model significantly ($P < .0001$) outperformed GBLUP, GEBLUP, and mmGBLUP models in PA across all scenarios. The increase in PA was particularly prominent when a trait had a higher h^2_{AE} or ρ_{AE}, indicating that mmGEBLUP exhibited a great advantage in predicting the complex trait of a target genotypes in a target environment. For example, when $\rho_A = 0.5$, $\rho_{AE} = 0.5$, $h^2_A = 0.3$, and h^2_{AE} increased from 0.2 to 0.3 and 0.6, the average PA increases of mmGEBLUP were 36.3%, 56.3%, and 103.3% compared to GBLUP; 27.9%, 36.4%, and 33.7% compared to GEBLUP; and 19.1%, 26.1%, and 25.8% compared to mmGBLUP (Fig. 3B). Similarly, when $\rho_A = 0.5$, $h^2_A = 0.3$, $h^2_{AE} = 0.3$, and ρ_{AE} increased from 0.3 to 0.5, the average PA remained the same at 0.502 for the mmGBLUP model but increased from 0.575 to 0.633 for the mmGEBLUP model (Fig. 3C). To conclude, these results demonstrate that h^2_A accounts for the baseline PA of the GBLUP model, h^2_{AE} is responsible for the accuracy increase between GBLUP and GEBLUP models, ρ_A explains the accuracy increase from the GEBLUP to the mmGBLUP model, and ρ_{AE} accounts for the accuracy increase from mmGBLUP to mmGEBLUP, as implied in Fig. 1. Thus, we believe that mmGEBLUP shows greater promise in capturing not only the genetic architecture of traits but also their GE interaction structures, from both the perspectives of heritability and polygenicity.

The value of the mmGEBLUP model extends beyond its higher PA; for instance, it facilitates the prediction and selection of superior varieties. The mmGEBLUP model assumes that individual GEBV can be assessed by four components, namely, the major additive BV (a^l), minor additive BV (a), major interaction BV (ae^l), and minor interaction BV (ae). The mmGEBLUP model predicts these four components for each individual in the validation environment. We compared the predicted values with the actual values in simulation (Fig. 3D–G). The correlation reached 0.686 and 0.625 for major and minor additive BVs and 0.799 and 0.291 for major and minor additive-by-environment BVs, respectively, demonstrating good accuracy of prediction. Based on the true or predicted values, we identified the top 20 individuals by their BVs, made up of four components (Fig. 3H and I). Among the top 20 individuals in their predicted values, 9 were also among the top 20 in their true values, including the top five individuals. These results underscore the effectiveness of the mmGEBLUP model in predicting BV and identifying superior lines for breeding programs.

Genetic mapping in rice and tobacco populations

We conducted genetic mapping in both rice and tobacco datasets (Figs S2 and S3), selected peak signals in a full genetic model including GE interaction, and compared the results with variance components analysis (Table 1). In the rice population, narrow-sense heritability estimates from peak variants aligned with the broad-sense heritability estimates from variance components analysis. Among the three traits, FLW exhibited the highest general and interaction heritabilities in the broad sense as well

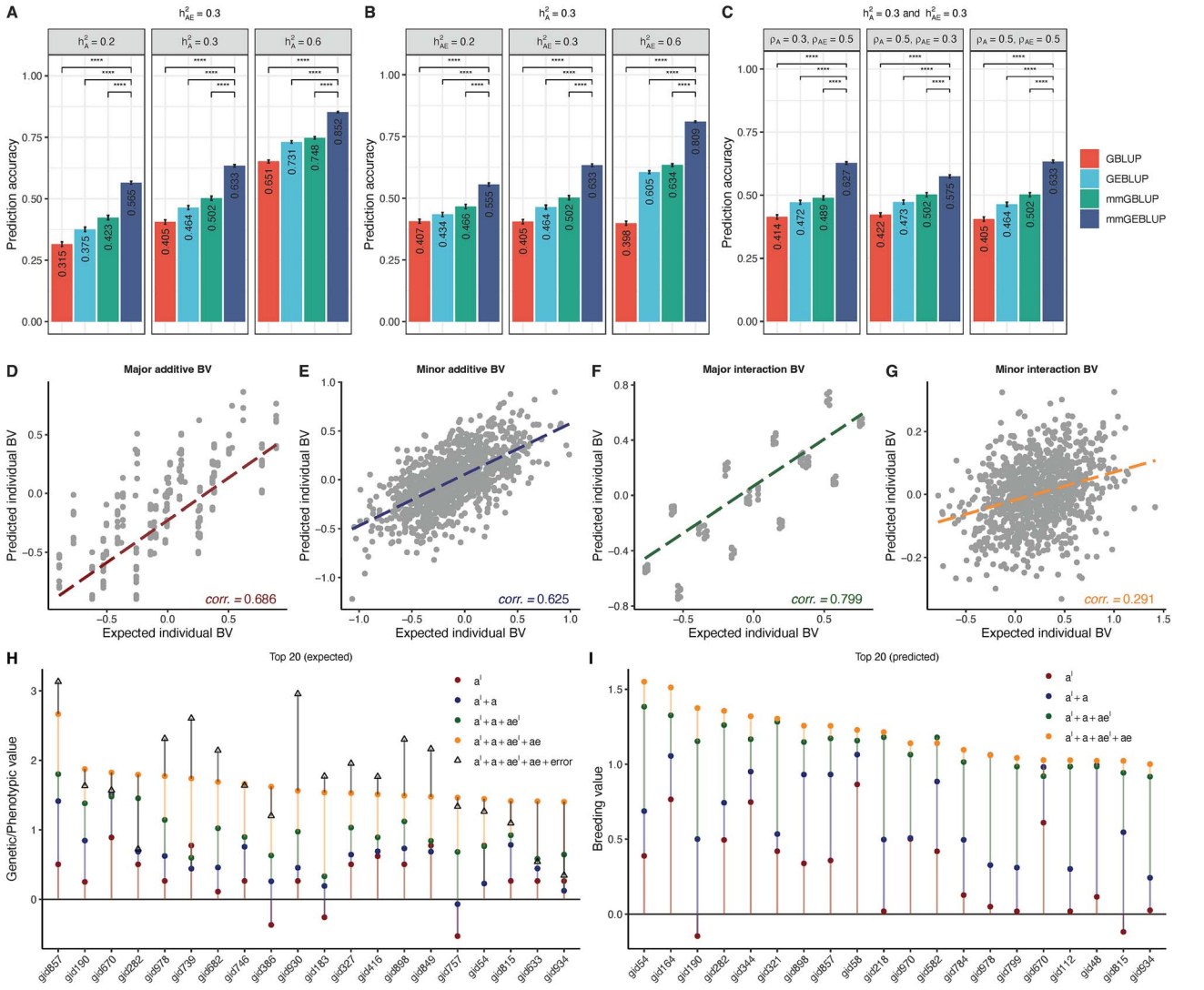

Figure 3. Performance of mmGEBLUP model in simulations. (A–C) Prediction accuracies of different simulated traits using GBLUP, GEBLUP, mmGBLUP, and mmGEBLUP models. Seven traits/scenarios are simulated to present different genetic architectures and GE interaction structures. In each scenario, the *P*-values for pair-wise Student's *t*-test are displayed on top of each plot in a 4-fold cross-validation with 10 repeats (df = 39, one-sided test). ****: $P \leq .0001$. Scenarios are organized in three groups: (A) traits with the same interaction heritability but different general heritability, $\rho_A = 0.5$, $\rho_{AE} = 0.5$, $h^2_{AE} = 0.3$, combined with three h^2_A (= 0.2, 0.3, and 0.6). (B) Traits with the same general heritability but different interaction heritability, $\rho_A = 0.5$, $\rho_{AE} = 0.5$, $h^2_A = 0.3$, combined with three h^2_{AE} (= 0.2, 0.3, and 0.6). (C) Traits with the same general and interaction heritability but different general and interaction polygenicity. (D–G) Regression plots of four expected and predicted BV components. (H, I) Top 20 individuals with their true and predicted four BV components. a^l: major additive BV, a: minor additive BV, ae^l: major interaction BV, ae: minor interaction BV, and error: true residual error. Individual BVs are given at $\rho_A = 0.5$, $\rho_{AE} = 0.5$, $h^2_{AE} = 0.3$, and $h^2_A = 0.3$.

as in the narrow sense, followed by PH in general heritability and YLD in interaction heritability. In the tobacco population, nLN had the highest general heritability (H^2_G = 48.5% and h^2_A = 41.9%), whereas the trait IL had the lowest general heritability (H^2_G = 4.3% and h^2_A = 9.3%) but the highest interaction heritability (H^2_{GE} = 63.0% and h^2_{AE} = 8.5%) among six traits. These findings, together with the following GP model evaluations and individual BV compositions, provide insights into the genetic architecture of traits.

Genomic prediction in rice population

We evaluated six experimental designs, as outlined in the original study [31], using the wet season of 2012 (2012WS) as the validation environment (Table S3). In Experiment design 2, up to seven environments were used as training environments, whereas Experiment design 7 included only the most recent two environments. The mmGEBLUP model significantly improved PA compared to mmGBLUP, GEBLUP, and GBLUP models across most experiments (Fig. 4A–C). On average, the mmGEBLUP model improved PA by 3.6%, 6.2%, and 6.0% for YLD; 1.0%, 1.7%, and 1.9% for PH; and 6.4%, 7.4%, and 8.1% for FLW, compared to mmGBLUP, GEBLUP, and GBLUP, respectively.

We further evaluated individual BVs predicted by the mmGEBLUP model across four components (Fig. 4D–F), using Experiment design 8 as an example. For all three traits, higher total BVs were primarily explained by major additive BVs. For FLW, the individual with the largest major additive BV ranked highest. However, for YLD, individuals with the largest major additive BVs, such as "A1271", "B1118", and "A1316" ranked lower due to negative minor BVs. When selecting environment-specific varieties, all four BV components should be considered, whereas for broadly adaptive varieties, comparing additive BVs alone is sufficient. Therefore, accessions like "A1271", "B1118", and "A1316" remain promising breeding materials due to their large additive BVs.

Table 1. Variance components and heritability estimation for traits in rice and tobacco populations.

Population	Trait[a]	Variance components Est. (S.E.)[b]				Heritability[c]			
		$\hat{\sigma}^2_G$	$\hat{\sigma}^2_E$	$\hat{\sigma}^2_{GE}$	$\hat{\sigma}^2_\varepsilon$	$\hat{H}^2_G$	$\hat{H}^2_{GE}$	$\hat{h}^2_A$	$\hat{h}^2_{AE}$
Rice	YLD	0.104 (0.018)	0.699 (0.077)	0.101 (0.016)	0.251 (0.018)	22.7%	22.2%	21.4%	9.6%
	PH	0.295 (0.018)	0.414 (0.086)	0.067 (0.017)	0.230 (0.019)	49.8%	11.3%	45.0%	7.7%
	FLW	0.450 (0.018)	0.281 (0.081)	0.187 (0.016)	0.093 (0.018)	61.7%	25.6%	70.6%	20.7%
Tobacco	nPH	0.467 (0.031)	0.143 (0.152)	0.337 (0.036)	0.177 (0.030)	47.6%	34.4%	24.1%	4.3%
	nLN	0.529 (0.032)	0.116 (0.153)	0.470 (0.037)	0.092 (0.030)	48.5%	43.1%	41.9%	1.8%
	SG	0.266 (0.031)	0.320 (0.153)	0.303 (0.035)	0.327 (0.030)	29.7%	33.8%	25.7%	0.0%
	IL	0.032 (0.029)	0.497 (0.148)	0.465 (0.035)	0.241 (0.030)	4.3%	63.0%	9.3%	8.5%
	LL	0.218 (0.032)	0.100 (0.153)	0.334 (0.037)	0.424 (0.031)	22.3%	34.2%	12.7%	0.0%
	LW	0.140 (0.032)	0.096 (0.152)	0.311 (0.036)	0.515 (0.031)	14.5%	32.2%	15.4%	0.6%

[a]Trait annotation: YLD, grain yield (kg/ha); PH, plant height (cm); FLW, flowering time (days to 50% flowering); nPH, natural plant height (cm); nLN, natural leaf number; SG, stem girth (cm); IL, internode length (cm); LL, length of the largest leaf (cm); and LW, width of the largest leaf (cm). [b]Variance components: Variance components analysis was performed based on the mixed linear model, $y = \mu + G + E + GE + \varepsilon$. Both the genetic effect (G), environment effect (E), GE interaction effect (GE), and residual effect (ε) are random effects. The *sommer* R package was employed to solve the mixed model equation and estimate the variance components. $\hat{\sigma}^2_G$: estimated genotypic variance; $\hat{\sigma}^2_E$: estimated environmental variance; $\hat{\sigma}^2_{GE}$: estimated genotype-by-environment interaction variance; $\hat{\sigma}^2_\varepsilon$: estimated residual variance. Numbers in brackets are standard errors (S.E.) of the variance component estimates. [c]Heritability: General heritability in the broad sense (H^2_G) and interaction heritability in the broad sense (H^2_{GE}) are estimated from variance components analysis, which are $\hat{H}^2_G = \hat{\sigma}^2_G / (\hat{\sigma}^2_G + \hat{\sigma}^2_{GE} + \hat{\sigma}^2_\varepsilon)$ and $\hat{H}^2_{GE} = \hat{\sigma}^2_{GE} / (\hat{\sigma}^2_G + \hat{\sigma}^2_{GE} + \hat{\sigma}^2_\varepsilon)$. General heritability in the narrow sense (h^2_A) and interaction heritability in the narrow sense (h^2_{AE}) are estimated based on peak variants detected by software QTXNetwork 2.0. $\hat{h}^2_A$ is calculated from the proportion of phenotypic variance explained by the additive effects of peak variants and $\hat{h}^2_{AE}$ is calculated from the proportion of phenotypic variance explained by the additive-by-environment interaction effects of peak variants.

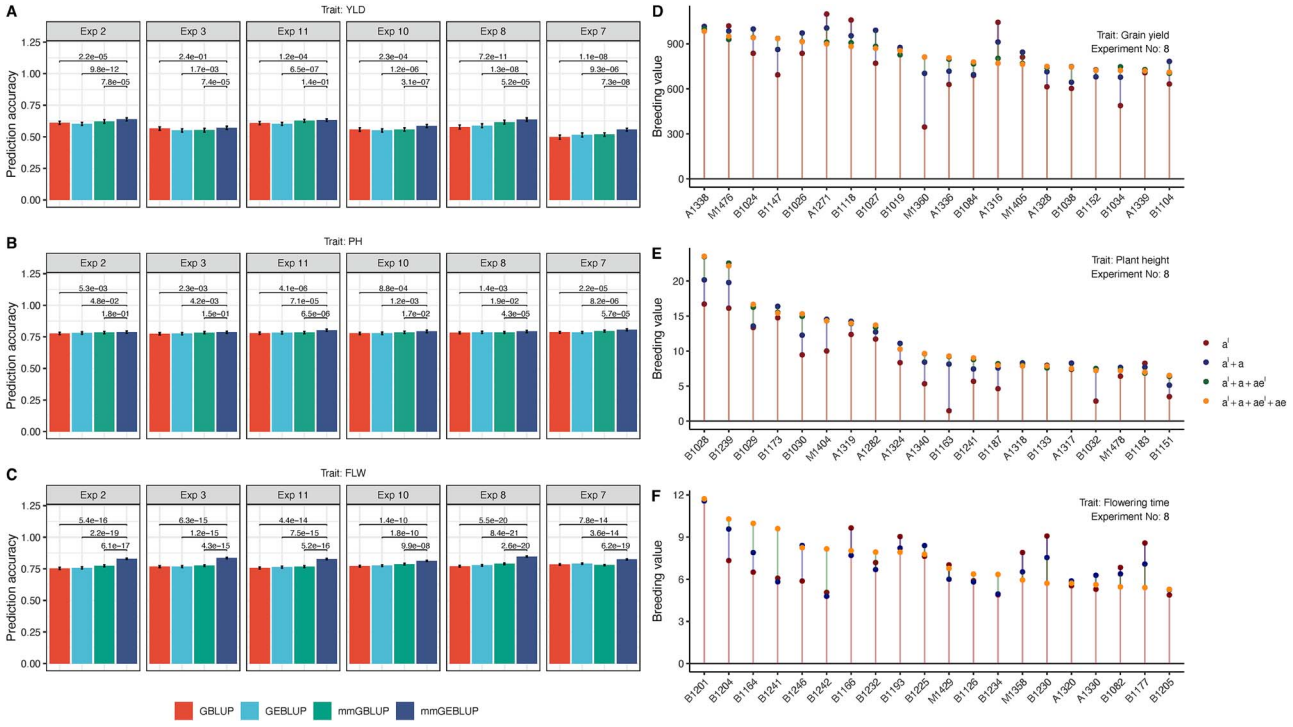

Figure 4. The performance of mmGEBLUP in the rice population. (A–C) Cross-validation prediction accuracies of YLD, PH, and FLW in the rice population using GBLUP, GEBLUP, mmGBLUP, and mmGEBLUP models. The *P*-values for pair-wise Student's *t*-test are displayed on top of each plot in a 5-fold cross-validation with 10 repeats (df = 49, one-sided). We include six experimental designs (Exp 2, Exp 3, Exp 11, Exp 10, Exp 8, and Exp 7) that consider 2012 WS as a validation environment. Experiment number is defined in the original study and is remade in Table S2. (D–F) Top 20 individuals with their four predicted BV components. a^l: major additive BV, a: minor additive BV, ae^l: major interaction BV, and ae: minor interaction BV.

A comprehensive interpretation of individual BVs helps breeders select elite accessions and refine breeding strategy. For example, MAS and GS can target major and minor additive BVs, respectively. Consider "B1201" and "B1204", the top two accessions with longer flowering times. While "B1201" had a larger major additive BV, its minor additive BV was smaller than that of "B1204". To potentially improve "B1204", MAS could be applied to its large-effect markers to resemble "B1201" in major genetic makeup while retaining its small-effect markers. This approach integrates the superiority of both "B1201" and "B1204" by pyramiding favorable major genes and background genes.

Genomic prediction in tobacco population

The evaluation of PA across models, particularly the relative increase in PA between models, emphasizes the significance of structural indicators (h^2_A, h^2_{AE}, ρ_A, and ρ_{AE}). Figure 5A illustrates the prediction performance of four GP methods for various agronomic traits, including the PA and pair-wise *P*-values from Student's

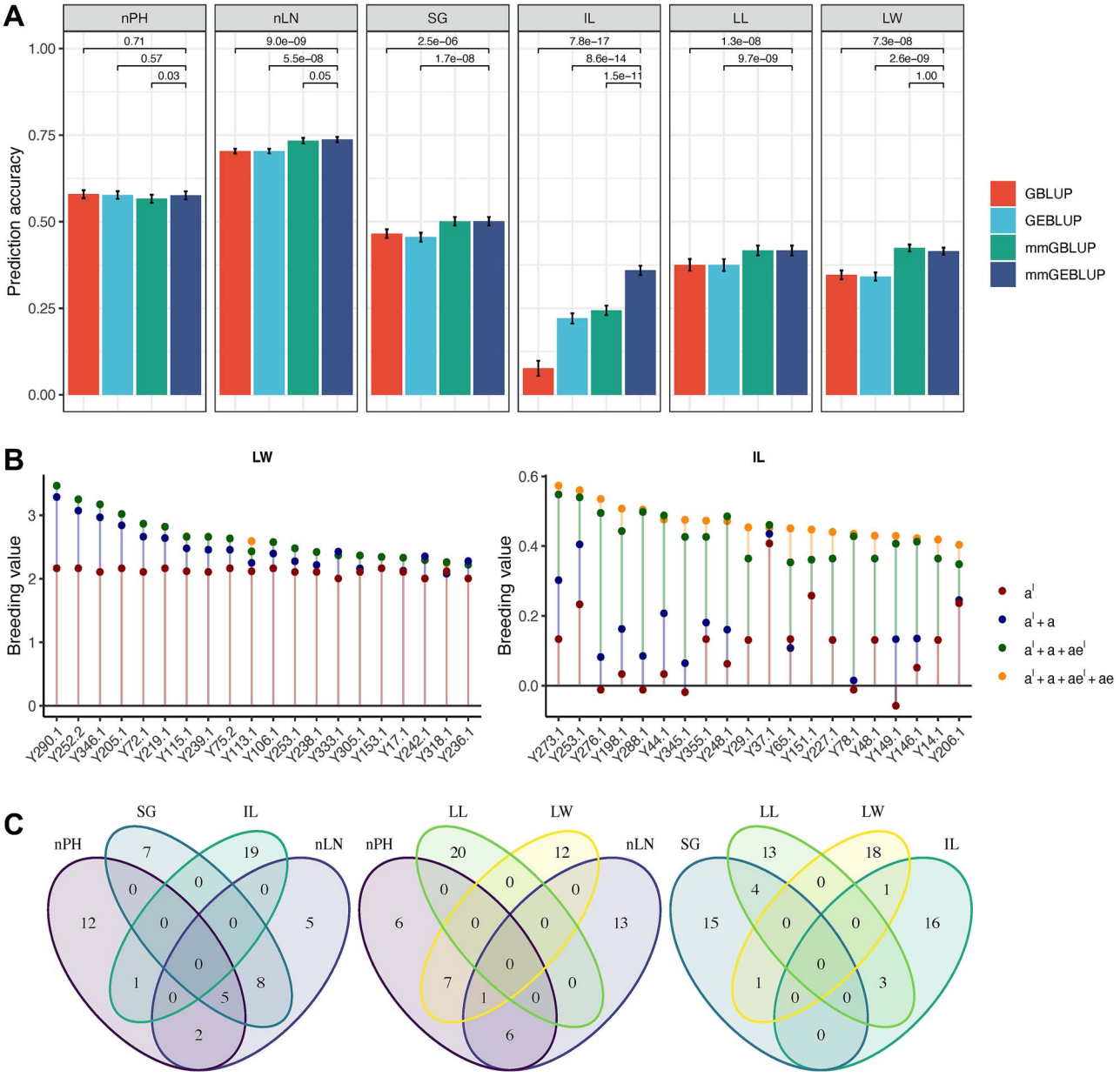

Figure 5. The performance of mmGEBLUP model in the tobacco population. (A) Cross-validation prediction accuracies of six agronomic traits in the tobacco population using GBLUP, GEBLUP, mmGBLUP, and mmGEBLUP models. The *P*-values for pair-wise Student's *t*-test are displayed on top of each plot in a 4-fold cross-validation with 10 repeats indicating (df = 39, one-sided). nPH, natural plant height; nLN, natural leaf number; SG, stem girth; IL, internode length; LL, length of the largest leaf; LW, width of the largest leaf. (B) Top 20 individuals with their predicted four BV components in traits LW and IL. a^l: major additive BV, a: minor additive BV, ae^l: major interaction BV, and ae: minor interaction BV. (C) Venn plots with top 20 individuals in six agronomic traits.

t-tests. The mmGEBLUP model significant improved PA for five traits ($P < 1e\text{-}5$) compared to GBLUP and GEBLUP and for three traits ($P < .05$) compared to the mmGBLUP model. Notably, IL exhibited the largest PA increase with mmGEBLUP: 49.6% compared to mmGBLUP, 69.5% compared to GEBLUP, and 114.7% compared to GBLUP.

The heritability estimates and PA comparisons across models provide insights into trait genetic architecture and GE interaction structures. Firstly, the relative performance of GBLUP reflects the relative magnitude of general heritability. For example, nLN, with the highest general heritability ($\hat{h}^2_A$ = 41.9%), showed the highest PA (0.704) with GBLUP, while IL, with the lowest heritability ($\hat{h}^2_A$ = 9.3%), exhibited the lowest PA (0.076) with GBLUP. Secondly, traits with high interaction heritability, such as IL, exhibited substantial PA increases with GEBLUP compared to GBLUP (e.g. PA increased from 0.076 to 0.221). Thirdly, mmGBLUP outperformed GEBLUP for traits with less polygenic architecture and larger ρ_A, as seen for LL and LW (PA increased from 0.375 to 0.417 for trait LL and from 0.341 to 0.424 for trait LW). Lastly, the significant PA difference between mmGEBLUP and mmGBLUP indicates a larger ρ_{AE}. For example, nPH, nLN, and IL showed *P*-values of .030, .049, and 1.50e-11, respectively. However, for SG and LL, where no GE interaction loci were detected, no difference was observed between mmGEBLUP and mmGBLUP.

Based on the mmGEBLUP model, individual BV predictions with four components are provided in Fig. 5B and Fig. S4. Interestingly,

for traits like nLN and LW, the major additive BVs for the top 20 individuals were substantially equal; it was the minor additive BV that caused the difference in the total BV. For example, "Y290.1" and "Y252.2" had identical major additive BVs (2.164), but their minor additive BVs were 1.123 and 0.908, respectively. For IL, both major additive and interaction BVs contributed to total BV variation. Major interaction BV should be considered for breeding environment-specific elite varieties. Certain accessions exhibited elite performance across multiple traits due to correlated traits and colocalized QTLs [9]. For example, five accessions, "Y287.1", "Y115.1", "Y360.1", "Y278.1", and "Y265.1" ranked among the top 20 for nPH, nLN, and SG (Fig. 5C).

Discussion

The genetic architecture of a trait is shaped by both major and minor genes, along with their core or peripheral interactions with the environment. Together, they guide primary breeding efforts. In this study, we present a new GP scheme that accurately models genetic architectures and provides practical suggestions based on four components of individual BV. Through extensive simulations, we show how stepwise increases in PA from GBLUP to GEBLUP, to mmGBLUP, and to mmGEBLUP indicate the scale of heritability and polygenicity in different traits. Applications to rice and tobacco traits showcased the scheme's ability to improve PA and predict individual BVs, emphasizing its value for breeding programs. Overall, mmGEBLUP stands out as a competitive and flexible GP model for multi-environment trials, excelling in capturing genetic architectures, predicting informative BVs, and aiding in the selection and improvement of complex crop traits.

While we have observed the robustness of the GBLUP model across various trait architectures, we believe in the benefits of developing statistical models that are tailored to particular traits. In practical applications, we encourage to explore different GP models as long as the computational cost is affordable. Beyond mixed linear models, the distinction between major and minor genetic variance can be incorporated into Bayesian frameworks. Nonparametric methods, such as machine learning and deep learning, further facilitate prediction without prior knowledge of potential genetic models [33].

Predicting individual BV is crucial for the genetic improvement of complex traits. GP emphasizes the concept of response to selection captured in the breeder's equation [34]. Traditional BV is defined as the additive genetic value while extending the concept to include additive-by-environment interaction is necessary for multi-environment data. In our model, we further decompose BV into major and minor components, for both additive and interaction effects. Both major and minor BVs play active roles in evaluation of individual performance. The relative size of major and minor additive BVs suggests the choice of performing MAS or GS or both. The major terms are easier to select and fix through MAS, while the minor terms act as an entirety in a collective whole, making them difficult to combine or transmit across varieties. Moreover, the proportion of GE interaction terms also gives information on how to breed environment-specific or broadly adaptive elite varieties. Ideally, the four predicted individual BVs can be used to design elite varieties by integrating desired elements from different individuals.

The genetic and GE interaction effects that we are currently focusing on are additive, using inbred lines. However, when breeding commercial lines with higher heterozygosity, dominance effects become non-negligible. The gap between broad-sense and narrow-sense heritability indicates the presence of nonadditive genetic variance (Table 1). Additionally, gene–gene (GG) interactions, particularly additive-by-additive interactions, play a crucial role in the genetic architecture of complex traits but were not included in our current analysis. This omission is basically due to the limited statistical power to detect GG interactions compared to GE interactions, given their small effect sizes and the large sample sizes required—empirical estimates show an average heritability contribution of only 0.06% per QTL pair for yield-related traits in rice [35]. Incorporating dominance and epistasis into our framework and expanding population sizes represent key directions for future research to further unravel the genetic architecture and improve the PA of the model.

Key Points

- We develop an advanced genomic prediction scheme by considering four genetic structural indicators, including newly applied concepts of general and interaction polygenicity, to predict complex traits with multi-environment trial data.
- Our systematic simulation and applications show the advantages of using the mmGEBLUP model, along with GBLUP, GEBLUP, and mmGBLUP models, by revealing the genetic architecture of the traits.
- We offer practical breeding guidance by predicting four components of breeding values, highlighting the contribution of minor effects in constructing total breeding values.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Conflict of interest: None declared.

Funding

This work was supported by the grants from the China National Tobacco Company (110202101038 (JY-15)), Yunnan Tobacco Company (2022530000241009), the Key Research and Development Program of Zhejiang Province (2022C02032), the Major Project of Natural Science Foundation of Zhejiang Province (LD25C130001), and the National Natural Science Foundation of China (31871707 and 31961143016).

Data availability

The following datasets can be downloaded from the links below:

The rice dataset: data are publicly available at https://datadryad.org/stash/dataset/doi:10.5061/dryad.7369p and https://datadryad.org/stash/dataset/doi:10.5061/dryad.vv28j

The tobacco dataset: data are deposited in the European Nucleotide Archive (ENA) repository (https://www.ebi.ac.uk/ena/browser/view/PRJEB59363).

The source code of R package "mmGEBLUP" is available at https://github.com/qixininin/mmGEBLUP.

Author contributions

Q.X.Z. and H.M.X. conceived the project; Q.X.Z., T.Z., and F.L. conducted data analyses; D.F., X.C., Z.T., and B.X. provided data and technical guidance; Q.X.Z., X.L., and H.M.X. wrote the manuscript;

All authors contributed to the manuscript and approved the submitted version.

References

1. Tian Z, Wang J-W, Li J. *et al.* Designing future crops: challenges and strategies for sustainable agriculture. *Plant J* 2021;**105**:1165–78. https://doi.org/10.1111/tpj.15107
2. Xu Y, Zhang X, Li H. *et al.* Smart breeding driven by big data, artificial intelligence, and integrated genomic-enviromic prediction. *Mol Plant* 2022;**15**:1664–95. https://doi.org/10.1016/j.molp.2022.09.001
3. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of Total genetic value using genome-wide dense marker maps. *Genetics* 2001;**157**:1819–29. https://doi.org/10.1093/genetics/157.4.1819
4. Begum H, Spindel JE, Lalusin A. *et al.* Genome-wide association mapping for yield and other agronomic traits in an elite breeding population of tropical Rice (Oryza sativa). *PloS One* 2015;**10**:e0119873. https://doi.org/10.1371/journal.pone.0119873
5. Famoso AN, Zhao K, Clark RT. *et al.* Genetic architecture of Aluminum tolerance in Rice (Oryza sativa) determined through genome-wide association analysis and QTL mapping. *PLoS Genet* 2011;**7**:e1002221. https://doi.org/10.1371/journal.pgen.1002221
6. Tian F, Bradbury PJ, Brown PJ. *et al.* Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat Genet* 2011;**43**:159–62. https://doi.org/10.1038/ng.746
7. Buckler ES, Holland JB, Bradbury PJ. *et al.* The genetic architecture of maize flowering time. *Science* 2009;**325**:714–8. https://doi.org/10.1126/science.1174276
8. Liu Y, Yuan G, Si H. *et al.* Identification of QTLs associated with agronomic traits in tobacco via a Biparental population and an eight-way MAGIC population. *Front Plant Sci* 2022;**13**. https://doi.org/10.3389/fpls.2022.878267.
9. Tong Z, Xu M, Zhang Q. *et al.* Construction of a high-density genetic map and dissection of genetic architecture of six agronomic traits in tobacco (Nicotiana tabacum L.). *Front Plant Sci* 2023;**14**:344. https://doi.org/10.3389/fpls.2023.1126529
10. Fragomeni BO, Lourenco DAL, Masuda Y. *et al.* Incorporation of causative quantitative trait nucleotides in single-step GBLUP. *Genetics Selection Evolution* 2017;**49**:59. https://doi.org/10.1186/s12711-017-0335-0
11. Bernardo R. Genomewide selection when major genes are known. *Crop Sci* 2014;**54**:68–75. https://doi.org/10.2135/cropsci2013.05.0315
12. Spindel JE, Begum H, Akdemir D. *et al.* Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. *Heredity* 2016;**116**:395–408. https://doi.org/10.1038/hdy.2015.113
13. Arruda MP, Lipka AE, Brown PJ. *et al.* Comparing genomic selection and marker-assisted selection for Fusarium head blight resistance in wheat (Triticum aestivum L.). *Mol Breeding* 2016;**36**:84. https://doi.org/10.1007/s11032-016-0508-5
14. Odilbekov F, Armoniené R, Koc A. *et al.* GWAS-assisted genomic prediction to predict resistance to Septoria Tritici blotch in Nordic winter wheat at seedling stage. *Front Genet* 2019;**10**. https://doi.org/10.3389/fgene.2019.01224.
15. Yin L, Zhang H, Zhou X. *et al.* KAML: improving genomic prediction accuracy of complex traits using machine learning determined parameters. *Genome Biol* 2020;**21**:1–22. https://doi.org/10.1186/s13059-020-02052-w
16. Brøndum RF, Su G, Janss L. *et al.* Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. *J Dairy Sci* 2015;**98**:4107–16. https://doi.org/10.3168/jds.2014-9005
17. Rice B, Lipka AE. Evaluation of RR-BLUP genomic selection models that incorporate peak genome-wide association study signals in maize and sorghum. *The Plant Genome* 2019;**12**:180052. https://doi.org/10.3835/plantgenome2018.07.0052
18. Zhang Y, Zhang M, Ye J. *et al.* Integrating genome-wide association study into genomic selection for the prediction of agronomic traits in rice (Oryza sativa L.). *Mol Breeding* 2023;**43**:81. https://doi.org/10.1007/s11032-023-01423-y
19. Xu Y, Zhang Y, Cui Y. *et al.* GA-GBLUP: leveraging the genetic algorithm to improve the predictability of genomic selection. *Brief Bioinform* 2024;**25**:bbae385. https://doi.org/10.1093/bib/bbae385
20. Chenu K. Chapter 13 - characterizing the crop environment – Nature, significance and applicationsIn: Sadras VO, Calderini DF, (eds.), *Crop Physiology*Second edn. San Diego: Academic Press, 2015, 321–48 10.1016/B978-0-12-417104-6.00013-3.
21. Burgueño J, de los Campos G, Weigel K. *et al.* Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. *Crop Sci* 2012;**52**:707–19. https://doi.org/10.2135/cropsci2011.06.0299
22. Schulz-Streeck T, Ogutu JO, Gordillo A. *et al.* Genomic selection allowing for marker-by-environment interaction. *Plant Breeding* 2013;**132**:532–8. https://doi.org/10.1111/pbr.12105
23. Heslot N, Yang HP, Sorrells ME. *et al.* Genomic selection in plant breeding: a comparison of models. *Crop Sci* 2012;**52**:146–60. https://doi.org/10.2135/cropsci2011.06.0297
24. Monteverde E, Rosas JE, Blanco P. *et al.* Multienvironment models increase prediction accuracy of complex traits in advanced breeding lines of Rice. *Crop Sci* 2018;**58**:1519–30. https://doi.org/10.2135/cropsci2017.09.0564
25. Acosta-Pech R, Crossa J, de los Campos G. *et al.* Genomic models with genotype × environment interaction for predicting hybrid performance: an application in maize hybrids. *Theor Appl Genet* 2017;**130**:1431–40. https://doi.org/10.1007/s00122-017-2898-0
26. Basnet BR, Crossa J, Dreisigacker S. *et al.* Hybrid wheat prediction using genomic, pedigree, and environmental Covariables interaction models. *The Plant Genome* 2019;**12**:180051. https://doi.org/10.3835/plantgenome2018.07.0051
27. Endelman JB, Jannink J-L. Shrinkage estimation of the realized relationship matrix. *G3 Genes|Genomes|Genetics* 2012;**2**:1405–13. https://doi.org/10.1534/g3.112.004259
28. Covarrubias-Pazaran G. Genome-assisted prediction of quantitative traits using the R package sommer. *PloS One* 2016;**11**:e0156744. https://doi.org/10.1371/journal.pone.0156744
29. Zhang FT, Zhu ZH, Tong XR. *et al.* Mixed linear model approaches of association mapping for complex traits based on omics variants. *Sci Rep* 2015;**5**:1–10.
30. Yang S, Zhou X. Accurate and scalable construction of polygenic scores in large biobank data sets. *The American Journal of Human Genetics* 2020;**106**:679–93. https://doi.org/10.1016/j.ajhg.2020.03.013
31. Spindel J, Begum H, Akdemir D. *et al.* Genomic selection and association mapping in Rice (Oryza sativa): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of Rice genomic selection in elite, tropical Rice breeding lines. *PLoS Genet* 2015;**11**:1–25. https://doi.org/10.1371/journal.pgen.1004982
32. Daetwyler HD, Villanueva B, Woolliams JA. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PloS One* 2008;**3**:e3395. https://doi.org/10.1371/journal.pone.0003395

33. Abdollahi-Arpanahi R, Gianola D, Peñagaricano F. Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. *Genetics Selection Evolution* 2020;**52**:12. https://doi.org/10.1186/s12711-020-00531-z
34. Cobb JN, Juma RU, Biswas PS. *et al.* Enhancing the rate of genetic gain in public-sector plant breeding programs: lessons from the breeder's equation. *Theor Appl Genet* 2019;**132**:627–45. https://doi.org/10.1007/s00122-019-03317-0
35. Wei X, Chen M, Zhang Q. *et al.* Genomic investigation of 18,421 lines reveals the genetic architecture of rice. *Science* 2024;**385**:eadm8762. https://doi.org/10.1126/science.adm8762

Briefings in Bioinformatics, 2025, **26(1)**, bbaf058
https://doi.org/10.1093/bib/bbaf058
Problem Solving Protocol