Tropical Plants

ARTICLE

Open Access

https://doi.org/10.48130/tp-0026-0023

Tropical Plants **2026**, 5: e027

Efficacy evaluation of a genomic selection breeding model for rice metabolites based on hyper-seq

Authors

Chengcai Xia[#], Xuan Luo[#], Yanjie Lu, Qi Liu, Hamada Mohamed Hassan, ..., Meiling Zou[*], Dayong Fan[*], Zhiqiang Xia[*]

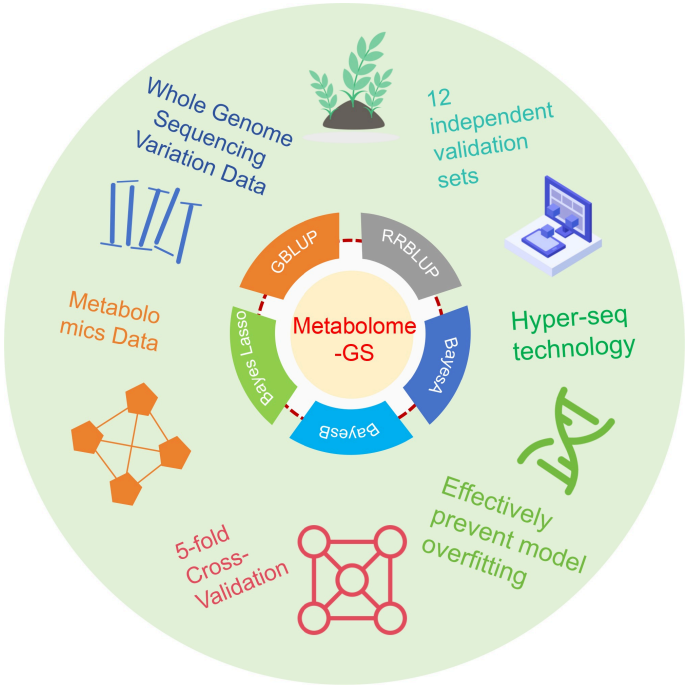
Correspondence

mlzou@hainanu.edu.cn (Zou M);
f3e902@outlook.com (Fan D);
zqiangx@gmail.com (Xia Z)

In Brief

We innovatively integrated metabolomic data into the genomic selection (GS) framework. Using a training population consisting of 524 diverse rice accessions and an independent validation set comprising 12 varieties, we evaluated the impacts of sequencing coverage depth (ranging from 4X to 30X) and genotype imputation on prediction accuracy.

Graphical abstract



Highlights

- The 12 validation sets are independent of the training cohort, effectively avoiding the risk of model overfitting.
- It is feasible to use metabolomic data for genomic selection (GS) prediction.
- The results indicate that low-coverage sequencing (4X) combined with imputation can achieve prediction accuracies comparable to full datasets.
- This work provides a theoretical and technical reference for integrating metabolomics into low-cost genomic breeding programs.

Citation: Xia C, Luo X, Lu Y, Liu Q, Hassan HM, et al. 2026. Efficacy evaluation of a genomic selection breeding model for rice metabolites based on hyper-seq. *Tropical Plants* 5: e027 https://doi.org/10.48130/tp-0026-0023

Tropical Plants

ARTICLE

Open Access

https://doi.org/10.48130/tp-0026-0023

Tropical Plants **2026**, 5: e027

Efficacy evaluation of a genomic selection breeding model for rice metabolites based on hyper-seq

Chengcai Xia[1#], Xuan Luo[1#], Yanjie Lu[1], Qi Liu[1], Hamada Mohamed Hassan[2], Isaev Vladislav[3], Meiling Zou[1*], Dayong Fan[4*] and Zhiqiang Xia[1*]

[1] *School of Breeding and Multiplication (Sanya Institute of Breeding and Multiplication)/School of Tropical Agriculture and Forestry/National Key Laboratory for Tropical Crop Breeding, Hainan University, Sanya 572025, China*

[2] *Rice Research Department - Sakha, Field Crops Research Institute, Agricultural Research Center, Ministry of Agriculture, Kafr El-Sheikh 33717, Egypt*

[3] *Shenzhen MSU-BIT University, Shenzhen 518172, China*

[4] *Guilin Agricultural Science Research Center, Guilin 541000, China*

[#] Authors contributed equally: Chengcai Xia, Xuan Luo

[*] Correspondence: mlzou@hainanu.edu.cn (Zou M); f3e902@outlook.com (Fan D); zqiangx@gmail.com (Xia Z)

Abstract

Plant metabolites play critical roles in regulating plant growth, development, and yield formation. Genomic selection (GS), as a cutting-edge breeding technology, holds significant promise for accelerating crop improvement and advancing precision breeding programs. Current GS models primarily focus on agronomic traits. In this study, we innovatively integrated metabolomic data into the GS framework to establish a multi-omics breeding approach. Leveraging publicly available genomic variation and metabolomic profiles from 524 diverse rice accessions, we systematically evaluated the prediction performance of five GS models. To further validate model applicability, we constructed genotype datasets for 12 core rice varieties using our independently developed Hyper-seq high-throughput sequencing platform, coupled with precise metabolic profiling. Crucially, this validation set remained independent of the training population, effectively mitigating overfitting risks. Our findings demonstrate for the first time the biological value of metabolomic data in genomic prediction. This work establishes a theoretical foundation and technical pathway for developing integrated multi-omics intelligence-driven breeding models.

Keywords: Rice, Metabolite, Genomic selection, *L*-Arginine, *N*-Feruloylputrescine

Citation: Xia C, Luo X, Lu Y, Liu Q, Hassan HM, et al. 2026. Efficacy evaluation of a genomic selection breeding model for rice metabolites based on hyper-seq. *Tropical Plants* 5: e027 https://doi.org/10.48130/tp-0026-0023

Introduction

As human society continues to evolve, breeding technologies have also been making continuous breakthroughs, with crop breeding progressing from the initial Breeding 1.0 era to the current Breeding 4.0 era. In the earliest Breeding 1.0 era, also known as the traditional breeding era, breeding was primarily based on breeders' observations and planting experiences[1]. In the early 20th century, breeding entered the 2.0 era, which was more scientific compared to the experience-driven 1.0 era. It employed strategies such as genetics and hybrid breeding for variety improvement. In the 3.0 era, molecular markers and gene editing breeding were widely applied in crop breeding. Currently, crop breeding is in the 4.0 era, characterized by intelligent breeding that integrates genotype- and phenotype-based technologies. It utilizes high-throughput sequencing technology and genomics, combined with integrated analysis of genomic, phenotypic, and environmental data, to enhance breeding efficiency and precision. Breeders can more accurately identify genes and molecular markers associated with specific traits, thereby accelerating the selection process for desirable traits[2]. Genomic Selection (GS) plays a crucial role in crop breeding today[3]. GS relies on genotyping technologies, utilizing high-density markers covering the entire genome for selective breeding. By combining phenotypic trait data with genotypic data, it establishes correlations between the two, marking the effects of all markers across the entire genome. Even if a single nucleotide polymorphism (SNP) has a small effect, it can still be captured and not overlooked due to its minor impact, leading to more accurate prediction results[4,5]. When genomic selection is combined with other technologies, it significantly shortens the generation interval, and prediction no longer depends on linkage disequilibrium between DNA markers and causative mutations. Since rapid breeding can substantially reduce the generation interval by applying genomic selection in each generation to screen parents for the next generation, the genetic gain achieved through this method can be greatly increased[6]. Currently, the cost of genotyping is the biggest challenge in implementing genomic selection. To reduce costs, one option is to apply genomic selection only every two or three generations or to select only candidate plants that have passed a reliable phenotypic analysis threshold for traits in rapid breeding cycles, such as certain disease-resistant types. In terms of sequencing technologies, low-coverage whole-genome sequencing (LcWGS) and microarray technologies are widely used in genomic selection breeding. LcWGS can be applied on a large scale for genotypic analysis of samples, obtaining high-density SNP markers at a relatively low cost[7]. However, due to its low coverage depth, LcWGS requires imputation and filling of missing genotypes to obtain more complete genotypic data[8,9]. Additionally, LcWGS demands high computational power because of its sequencing depth limitations, requiring substantial computational support to complete genotype imputation. Using imputed genotypic data and existing phenotypic data, prediction models are constructed to estimate the genomic estimated breeding values (GEBVs) of individuals. Apart from LcWGS, microarrays are also widely used in GS breeding. Different

microarrays can be designed based on various needs, but when analyzing a large number of samples, it can incur high costs. Hyper-seq is a low-cost, efficient, flexible, and high-throughput DNA sequencing library preparation and genotyping method that significantly reduces the time and cost of library construction and sequencing[10]. This method not only obtains markers across the entire genome but also greatly reduces the cost and time of library construction, improves efficiency, and yields accurate results.

Rice (*Oryza sativa*) is one of the most crucial food crops globally, playing a vital role in ensuring global food security. China has made remarkable achievements in rice breeding, cultivation management, and yield enhancement, contributing significantly to the global food supply. In the realm of GS breeding, current research primarily focuses on key rice traits, such as plant height, heading date, and grain quality. Xiao et al. conducted resequencing on 200 japonica rice varieties and identified 2,410,743 SNPs. They utilized Ridge Regression Best Linear Unbiased Prediction (RRBLUP) to predict traits including yield per plant, number of panicles per plant, number of grains per panicle, and thousand-grain weight, with prediction accuracies ranging from 0.52 to 0.917. They found that the actual yield-related trait values of YG7313 and NG9108 were highly consistent with their predicted values. Consequently, using these two varieties as parents for variety breeding, they successfully developed two backbone lines, XY99 and JXY1, strongly demonstrating the practicality of GS breeding in rice[11]. Cui et al. collected published data and established three different populations for GS breeding research, employing the Best Linear Unbiased Prediction (BLUP) model. Initially, they analyzed the prediction accuracy of multiple important traits in a population of 1,495 rice hybrid combinations, finding that the prediction accuracy for seven traits was all above 0.6. Subsequently, they predicted 4,498,500 potential combinations from 3,000 rice materials. The research results indicate that GS holds great promise in hybrid breeding, capable of saving substantial costs[12]. Xu et al. utilized metabolomic data to predict hybrid yield in rice. This study compared six GS prediction methods, including Lasso, BLUP, Partial Least Squares (PLS), and Bayesian Generalized Linear Regression (BGLR), for predicting values from genomic, transcriptomic, and metabolomic data. The results revealed that Lasso and BLUP were the most effective methods, and that metabolomic data were more effective than genomic data in predicting yield; in fact, the predictability of hybrid yield nearly doubled[13].

Plant metabolites play a pivotal role in maintaining crop yield and nutrient content, which are critical for global food security. Currently, there is relatively limited research on genomic selection breeding targeting metabolite traits, yet the significance of metabolites in breeding research cannot be overstated. In this study, we utilized rice resequencing data and metabolomic data collected from the National Key Laboratory of Crop Genetic Improvement and the National Center of Plant Gene Research (Wuhan, China)[14]. By integrating high-throughput sequencing technology, Hyper-seq, we conducted genomic selection breeding for rice metabolites. We employed five genomic selection models—GBLUP, RRBLUP, Bayes A, Bayes B, and Bayes Lasso—to evaluate the prediction accuracy of these metabolite traits. In this investigation, we focused on two metabolites: *L*-Arginine and *N*-Feruloylputrescine. *N*-Feruloylputrescine is a plant secondary metabolite belonging to the phenolamide class, which plays a crucial role in defending rice against herbivorous damage[15]. L-Arginine metabolism is closely linked to nitrogen uptake, storage, and cycling in rice. Moreover, arginine is essential for the normal growth of rice roots[16]. Exogenous application of arginine has been shown to have the most pronounced effect in promoting the accumulation of photosynthetic products in rice seedling leaves, facilitating the efficient conversion of inorganic carbon into the organic carbohydrates required by seedlings[17]. This study aimed to explore the application of genomic selection in rice metabolite traits by focusing on these two metabolites. By integrating resequencing data and Hyper-seq data, we sought to accelerate rice breeding, shorten the breeding cycle, reduce breeding costs, and provide valuable insights for subsequent related research.

It is worth noting that early genomic selection (GS) solely relied on genotype (G) information for phenotypic prediction[18]. In recent years, a substantial number of studies have incorporated environmental factors (E) into the model (G×E), significantly enhancing predictive ability[19]. However, the high cost of environmental data collection, low repeatability, and the difficulty in standardizing the climate-soil combinations across different experimental stations have hindered its immediate application in large-scale breeding programs[20]. Metabolites, serving as an 'intermediate layer between G and E,' are regulated by genotype while also responding rapidly to environmental fluctuations[21]. They can be regarded as 'genotype-metabolite' (G × M) covariates that indirectly carry G × E information[22]. Therefore, this study aims to explore the application value of metabolite data by integrating metabolic data and considering the influence of genotype-metabolite (G × M) interactions, thereby expanding the data sources and generalization scope of genomic selection (GS) breeding and constructing more robust prediction methods. Although the phenotypic data currently originate solely from leaf samples, we have employed a dual strategy of 5-fold cross-validation combined with independent validation to establish the baseline predictive power of G × M. Subsequently, we will explicitly estimate the relative contributions of G × E and G × M through multi-location trials, gradually extending metabolite-assisted GS to real-world breeding scenarios[23].

Materials and methods

Data collection

We obtained genomic and metabolomic data for 524 core rice germplasm resources from the RiceVarMap2 public data platform. For validation purposes, 12 rice materials were provided by the Guilin Agricultural Science Research Center, and library construction and sequencing were performed using the Hyper-seq technology. Meanwhile, we conducted metabolite detection on the leaves of these 12 validation samples. After freeze-drying the leaves, they were ground into powder. Subsequently, 25 mg of each sample was accurately weighed at low temperature into an EP tube, along with homogenization beads. Then, 1,000 μL of extraction solution (methanol : acetonitrile : water = 2:2:1 (V/V), containing a mixture of isotopically labeled internal standards) was added. A Vanquish (Thermo Fisher Scientific) ultra-high-performance liquid chromatograph equipped with a Phenomenex Kinetex C18 (2.1 mm × 50 mm, 2.6 μm) liquid chromatography column was used for the chromatographic separation of target compounds.The mobile phase A is an aqueous phase containing 0.01% acetic acid, and the mobile phase B is a mixture of isopropanol and acetonitrile (1:1, v/v). The sample tray temperature is set at 4 °C, and the injection volume is 2 μL. The Orbitrap Exploris 120 mass spectrometer can perform first-order and second-order mass spectrometry data acquisition under the control of the control software (Xcalibur, version 4.4, Thermo). The detailed parameters are as follows: Sheath gas flow rate: 50 Arb; Aux gas flow rate: 15 Arb; Capillary temperature: 320 °C; Full ms

resolution: 60,000; MS/MS resolution: 15,000; Collision energy: Staggered Normalized Collision Energy (SNCE) 20/30/40; Spray Voltage: 3.8 kV in positive mode or −3.4 kV in negative mode.

Dataset processing for model training and validation

In this study, ten types of genotype datasets were systematically constructed to evaluate the influence of sequencing coverage and genotype interpolation on the prediction accuracy. We initially filtered the genotype data of 524 rice germplasm resources using vcftools with the parameters (--minQ 30 --minDP 3 --max-missing 0.9). Subsequently, we employed bamdst (version 0.3.5) and bedtools (version 2.30.0) to extract the overlapping regions between the genotype data of 12 validation samples and the filtered genotype data of the 524 rice germplasm resources. Based on these overlapping regions, we constructed the baseline dataset ALL. Next, we performed genotype imputation on the ALL dataset using Beagle (version 5.4). During the imputation process, we set the default window size to 50 Mb and the number of iterations to ten, ultimately generating the imputed dataset ALLIP. In the coverage-based stratified sampling step, we screened the original dataset according to different sequence depth thresholds. Specifically, we extracted single-nucleotide polymorphism (SNP) sites with coverage levels of ≥ 4X, ≥ 10X, ≥ 15X, and ≥ 30X from the ALL dataset, and constructed the 4X, 10X, 15X, and 30X datasets, respectively. Meanwhile, we extracted sites with the same coverage thresholds from the ALLIP dataset to construct the 4XIP, 10XIP, 15XIP, and 30XIP datasets. Finally, we successfully obtained ten datasets, namely ALL, 4X - 30X, and ALLIP, 4XIP - 30XIP. The coverage distributions of these datasets exhibited a distinct stepwise pattern.

GS model training and validation

In this study, five genomic selection models, namely GBLUP, RRBLUP, BayesA, BayesB, and Bayes Lasso (https://easygene.console.aliyun.com/workspaces), were employed to investigate their accuracy in predicting rice metabolites. We utilized a five-fold cross-validation approach to evaluate the prediction accuracy. Specifically, the 524 rice samples were divided into five subsets, with 80% of the samples selected as the training population and the remaining 20% as the prediction population in each iteration. Subsequently, we calculated the correlation between the observed and predicted values. To ensure the reliability of the results, this process was repeated five times, and the average of the five prediction correlation coefficients was computed. Additionally, we validated the prediction accuracy of each model using 12 rice samples that were independent of the training population, thereby mitigating the risk of model overfitting.

Results

Genotypic data and metabolic phenotypic data analysis

Genotype data analysis

After aligning the sequencing data of 12 rice samples constructed using the Hyper-seq library with the rice reference genome, it was found that the proportion of high - quality coverage regions (coverage depth > 10) in the rice genome was approximately 38%. We selected these high-quality coverage regions as the Hyper-seq coverage regions for rice flour rice. Subsequently, high-quality SNPs were extracted from the 524 rice re-sequencing data located within the Hyper-seq coverage regions of rice flour rice, resulting in a total of 38,016 high-quality SNP sites in the Hyper-seq coverage regions. Then, genotype imputation was performed on the genotype data of 524 rice samples and the Hyper-seq data to obtain the imputed genotypes. For the genotype data of 524 rice samples used in the model training and prediction sets, SNP site imputation was carried out using Beagle 5.4 software. After imputation, the missing rate of genome-wide SNP sites decreased from 0.34% in the original data to 0. By setting read coverage thresholds (≥ 4X, 10X, 15X, and all sites), the numbers of effective SNPs in the population were 7,566 (4X), 2,276 (10X), 1,448 (15X), 622 (30X), and 38,016 (ALL).

Phenotypic data analysis

By comparing the metabolomic detection data of 524 training and prediction samples with those of 12 validation samples, we identified 15 metabolites that could be stably detected in both groups of samples. These metabolites included three flavonoids, four alkaloids, two nucleosides, two small peptides, one lipid, one coumarin, as well as trigonelline and choline. Statistical analysis revealed that the coefficients of variation (CV) of the contents of these 15 metabolites in the training population ranged from 1.91% to 9.98%. After conducting density distribution analysis on the contents of these 15 metabolites in 524 rice samples, we found that the content of each metabolite did not significantly deviate from a normal distribution in the 524 samples (Fig. 1): from the analysis of metabolite contents, in terms of distribution characteristics and based on relevant density distribution analysis, most metabolite contents exhibit a relatively regular distribution pattern and approximately conform to a normal distribution. This indicates that metabolite contents have a certain central tendency and symmetry within the sample population, with a relatively concentrated value range. The metabolite contents of most samples fluctuate around the mean. This result implies that the content distribution of these metabolites in rice samples possesses a certain degree of regularity and stability, providing a reliable data foundation for subsequent research on rice metabolites based on the normal distribution assumption. Based on the kernel density estimation curve and the standard normal distribution curve, we calculated the mean squared error (MSE) of the content of each metabolite. The results showed that the MSEs of the contents of these 15 metabolites ranged from 0.00019 to 0.00374 (Table 1). These findings indicate that the contents of the 15 selected metabolites vary among the 524 samples and can be used for the initial training and prediction of models.

Metabolite-genotype joint model training based on SNP coverage optimization

During the training of five genomic selection models, we utilized both unimputed and imputed datasets. When training with the complete unimputed dataset, the prediction accuracy for different traits and models ranged from 5.81% to 54.21%. Among them, Trigonelline exhibited the highest prediction accuracy (with an average of 52.97%), while 1-Monopalmitin showed the lowest (with an average of 12.17%). Subsequently, we extracted SNP loci with coverage depths greater than 4X, 10X, 15X, and 30X for model prediction. We observed that, although the number of SNP loci with coverage greater than 4X accounted for only one-fifth of the total SNPs, their prediction accuracy was comparable to that achieved using all SNPs for model training, and in some cases, even showed an upward trend across different traits and models. When using SNP loci with coverage greater than 10X, 15X, and 30X for model training, we

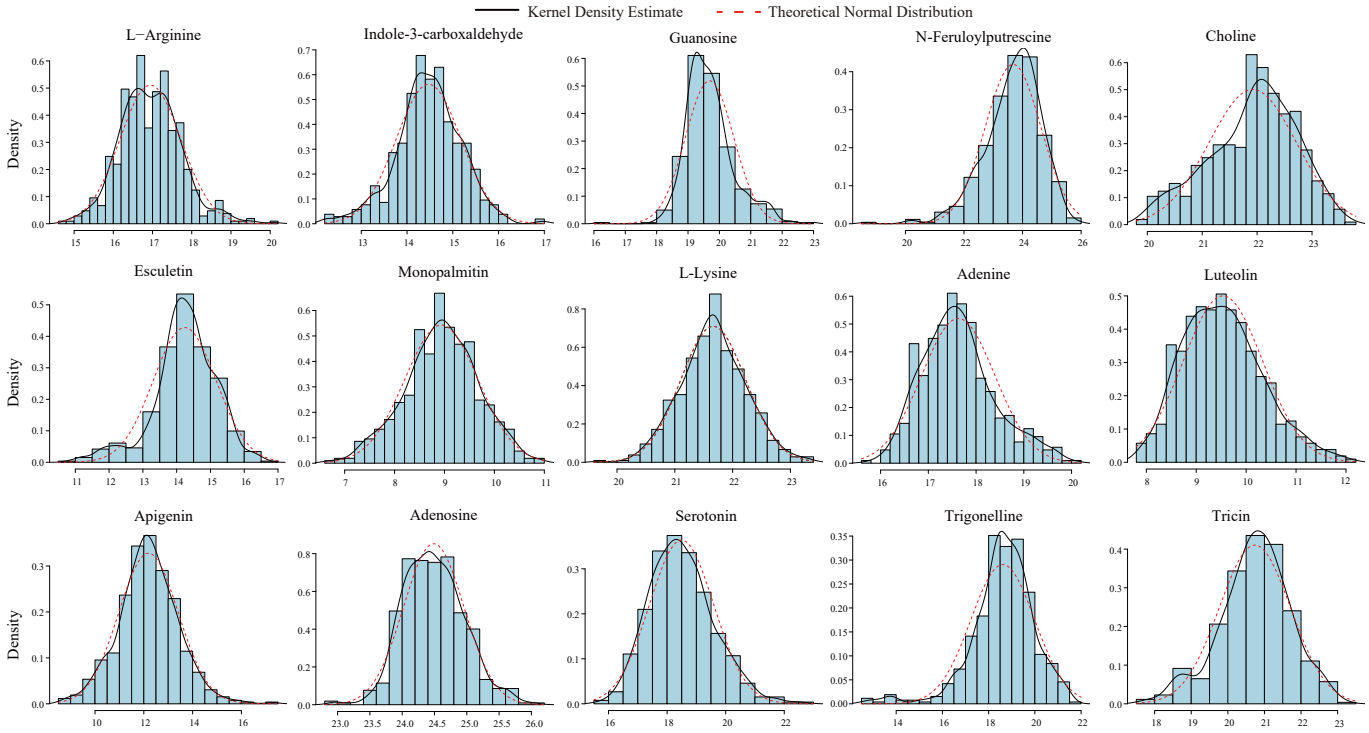


Fig. 1 Distribution characteristics of metabolite content.

Table 1. Descriptive statistical analysis of each metabolite.

	MSE	ISE	CV (%)
Apigenin	0.00019	0.0016	9.98
1-Monopalmitin	0.00033	0.0014	8.22
Serotonin	0.00039	0.0027	5.93
L-Lysine	0.00050	0.0018	2.60
L-Arginine	0.00053	0.0029	4.61
Tricin	0.00070	0.0040	4.69
Trigonelline	0.00089	0.0081	7.38
Indole-3-carboxaldehyde	0.00089	0.0043	4.91
N-Feruloylputrescine	0.00114	0.0083	4.02
Luteolin	0.00120	0.0050	8.37
Adenosine	0.00192	0.0059	1.91
Adenine	0.00216	0.0094	4.35
Esculetin	0.00223	0.0130	6.55
Choline	0.00278	0.0109	3.63
Guanosine	0.00374	0.0252	3.90

MSE, mean squared error; ISE, integral square error; CV, coefficient of variation.

analyzed the prediction accuracy results and calculated the standard deviations across different metabolites. The results indicated that for two alkaloids and lipid metabolites, the prediction accuracy began to decline significantly when using SNP loci with coverage ≥15X, and the standard deviations also exhibited high values. Notably, the GBLUP model showed a considerable deviation from the overall performance level for Adenine and 1-Monopalmitin. Even under these circumstances, it was evident that when using SNP loci with coverage ≥ 4X for training, the prediction accuracy remained close to, or even surpassed, that achieved using all SNPs. *N*-Feruloylputrescine displayed the highest standard deviation (7.55%) across various models, while Serotonin showed the lowest (0.93%). The prediction accuracy for *L*-Lysine, *L*-Arginine, and Trigonelline was all above 40%, with relatively small standard deviations of 2.67%, 2.37%, and 3.27%, respectively. Subsequently, we imputed all genotypic data and continued with model training. The results showed that the three metabolites with the highest prediction accuracy in the unimputed dataset (*L*-Lysine, *L*-Arginine, Trigonelline) maintained high accuracy in the imputed dataset, but their standard deviations decreased from the original 2.67%, 2.37%, and 3.27% to 2.07%, 1.90%, and 4.11%, respectively. Moreover, the overall prediction accuracy across all metabolites and models did not significantly change after imputation. These findings suggest that the low-coverage (4X) model can achieve prediction accuracy comparable to that of the full-locus model, even with a substantial reduction in the number of SNPs (only 20.1% of the full-locus data). Therefore, when conducting GS breeding on a large scale, it is feasible to appropriately reduce the sequencing coverage depth to lower costs while maintaining prediction accuracy.

The prediction accuracy of the independent validation set model

Subsequently, we utilized 12 rice germplasm resources as an independent validation set to specifically evaluate the performance of all the aforementioned models for *L*-Arginine and *N*-Feruloylputrescine. *L*-Arginine achieved the highest prediction accuracy using the RRBLUP and GBLUP models, with no significant change in the Pearson correlation coefficient between the predicted and measured values compared to the training set (Fig. 2a). Similarly, the predictive performance of models trained with imputed genotypic data remained comparable to that of models trained with unimputed data (Fig. 2b). For *N*-Feruloylputrescine, the Bayes B model demonstrated the best performance (Fig. 2c, d). Notably, during the prediction process for *N*-Feruloylputrescine in the validation set, the metabolite exhibited greater sensitivity to the number of SNP loci compared to *L*-Arginine, with a noticeable decline in prediction accuracy as the number of SNP markers decreased. By calculating the standard deviations of the prediction results across five models using all SNPs, as well as SNPs with coverage depths of 4X, 10X, 15X, and 30X, we observed that, regardless of whether the genotypic

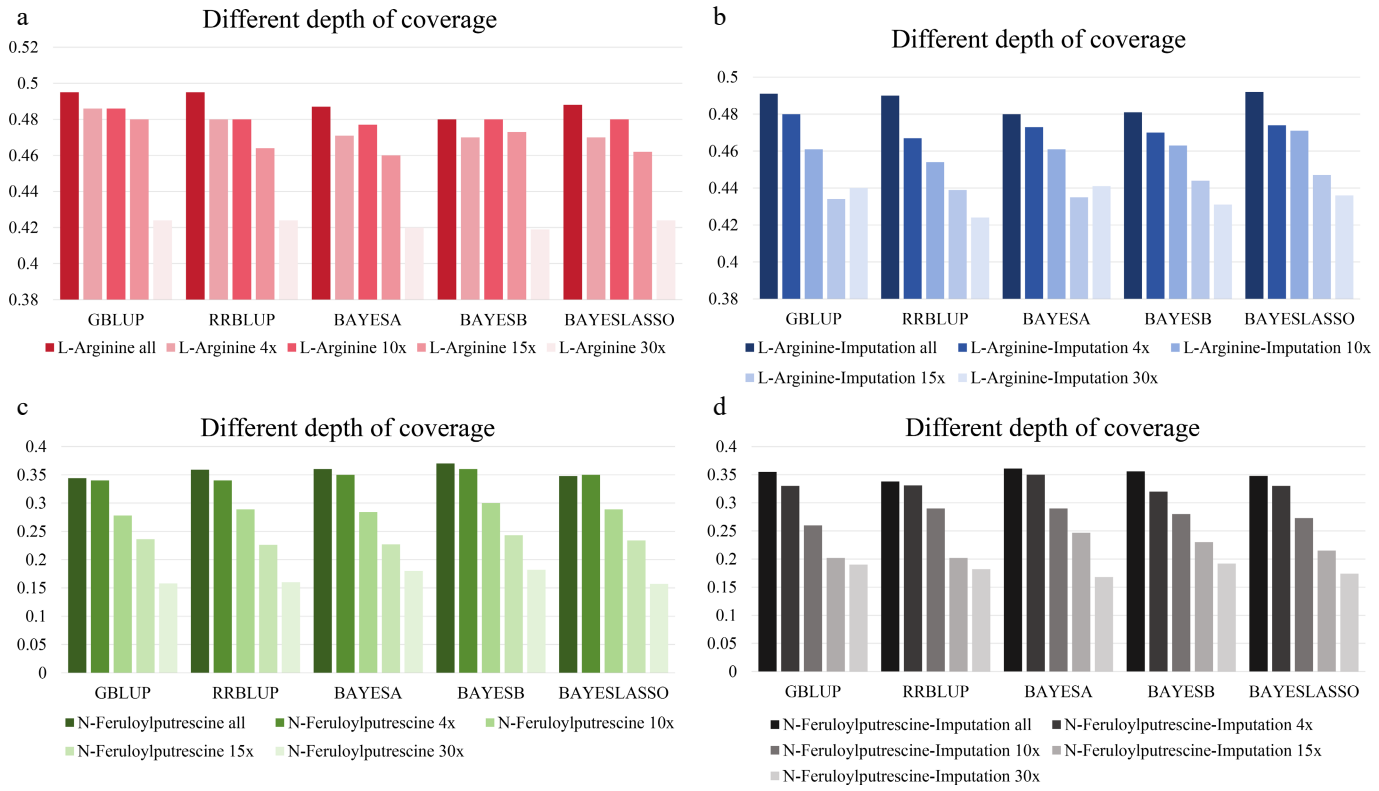


Fig. 2 The prediction accuracy results of the independent validation set under different models. (a) The prediction accuracy of *L*-Arginine in the model trained on the genotype data before interpolation. (b) The prediction accuracy of *L*-Arginine in the model trained on the genotype data for interpolation. (c) The prediction accuracy of *N*-Feruloylputrescine in the model trained on the genotype data before interpolation. (d) The prediction accuracy of *N*-Feruloylputrescine in the model trained on the genotype data for interpolation.

data were imputed or not, the standard deviations obtained for *N*-Feruloylputrescine were consistently higher than those for *L*-Arginine. These findings confirm that integrating metabolomic data can enhance the generalization ability of genomic selection (GS) models.

Discussion

Genomic selection (GS), as a revolutionary breeding technology, is rapidly transforming the landscape of traditional breeding. This technique relies on high-density molecular markers spanning the entire genome and integrates phenotypic data to enable breeders to estimate the breeding values of individuals. Consequently, it allows for precise selection and rapid aggregation of desirable traits. In recent years, with the advancement of high-throughput sequencing technologies and the reduction in sequencing costs, GS technology has also made significant strides in crop breeding. Particularly in hybrid crop breeding, where the genotypes of hybrid offspring can be inferred from those of their parents, the advantages of GS become even more pronounced. Currently, GS validation studies have been conducted on various crops both domestically and internationally[24,25].

The continuous development of SNP (Single Nucleotide Polymorphism) breeding chips for crops has provided crucial technological support for obtaining genotype data in Genomic Selection (GS). Currently, over a hundred chips have been developed for more than 25 crop species. Therefore, reducing the cost of individual genotyping has remained a research focus in GS[26,27]. Although gene chip technology is relatively cost-effective compared to high-throughput sequencing, the cost of gene chips is still high, especially for large-scale applications. Different types of gene chips need to be designed and produced for specific research purposes, which increases the complexity and cost of chip preparation. Moreover, the SNP site density of gene chips may not be sufficient to cover all genetic variations. Gene chips require prior knowledge of the genome-wide SNP information of a species (typically obtained from large-scale resequencing) to select SNPs for chip design. Consequently, their ability to detect new variations is limited[28]. Hyper-seq technology, with its more comprehensive sequencing scope, remarkable stability, and proven reliability in genomic selection prediction across multiple species, overcomes the limitations of gene chips and low-coverage resequencing. By integrating with traditional models, it significantly reduces breeding costs, enhances the selection efficiency of key agronomic traits, injects impetus into the development of intelligent breeding 4.0, and propels the breeding industry toward a more precise and efficient new stage[7,9,29–32].

Rice (Oryza sativa L.) is one of the most important food crops globally. As a monocotyledonous model plant, research on its rich metabolome has made significant contributions to rice breeding practices. However, there is currently a substantial gap in genomics-based selection breeding targeting metabolites. Most studies have focused on rice traits such as yield, plant height, grain length, heading date[24,33,34], rice grain appearance quality, and disease resistance[35,36]. In this study, by integrating high-density SNP data with metabolomics analysis, we systematically evaluated the impact of different sequencing depths on the prediction models of rice metabolites, revealing a complex interplay between coverage optimization and metabolite characteristics. Metabolites, as products of

biochemical reactions within organisms, directly reflect the outcomes of genotype-environment interactions. During growth and development, plants produce a vast array of metabolites to adapt to changing environments and various stresses. Both internal and external factors can lead to changes in specific metabolites[37]. This study focuses on 15 metabolites. As can be observed from Fig. 3, prior to data imputation (taking low-coverage data as an example), these 15 metabolites exhibited extremely significant differences in performance across different models, with substantial data fluctuations. Specifically, under different coverage levels or data combination scenarios, the percentage values corresponding to each metabolite demonstrated a high degree of dispersion. For some metabolites (in the context of relevant model results), the percentages were close to 0%, while for others, they exceeded 50%. This fully indicates that without data imputation, the data quality was highly unstable. The data of different metabolites were constrained by the conditions of the original data, showing great uncertainty under various scenarios, which consequently led to marked differences in the results of models constructed based on such data.

After data imputation, the data exhibited markedly different characteristics. The model results corresponding to the 15 metabolites were more stable compared to those before imputation, with a more concentrated distribution of values. Although there were still some fluctuations, the overall data variability decreased significantly. This suggests that during the data processing, the imputation method played a certain role in standardizing and optimizing the data of the 15 metabolites, effectively reducing the interference caused by the instability of the original data on the model results. By comparing the prediction accuracy of datasets under various models, we found that the prediction accuracy for some metabolites using a subset of SNP loci even surpassed that of the full-locus model. Specifically, the accuracy for flavonoids (*L*-Lysine, *L*-Arginine) reached 40.2% in the 4X group (a 1.1% improvement over the full-locus model), and after Beagle imputation, the standard deviation was further reduced by 32% (e.g., for alkaloid m, it decreased from 3.27% to 2.05%). Simple primary metabolites like *L*-Arginine (Fig. 2a, b) exhibited remarkable stability in the RRBLUP/GBLUP models, whereas complex secondary metabolites like *N*-Feruloylputrescine (Fig. 2c, d) showed a sharp decline in accuracy.

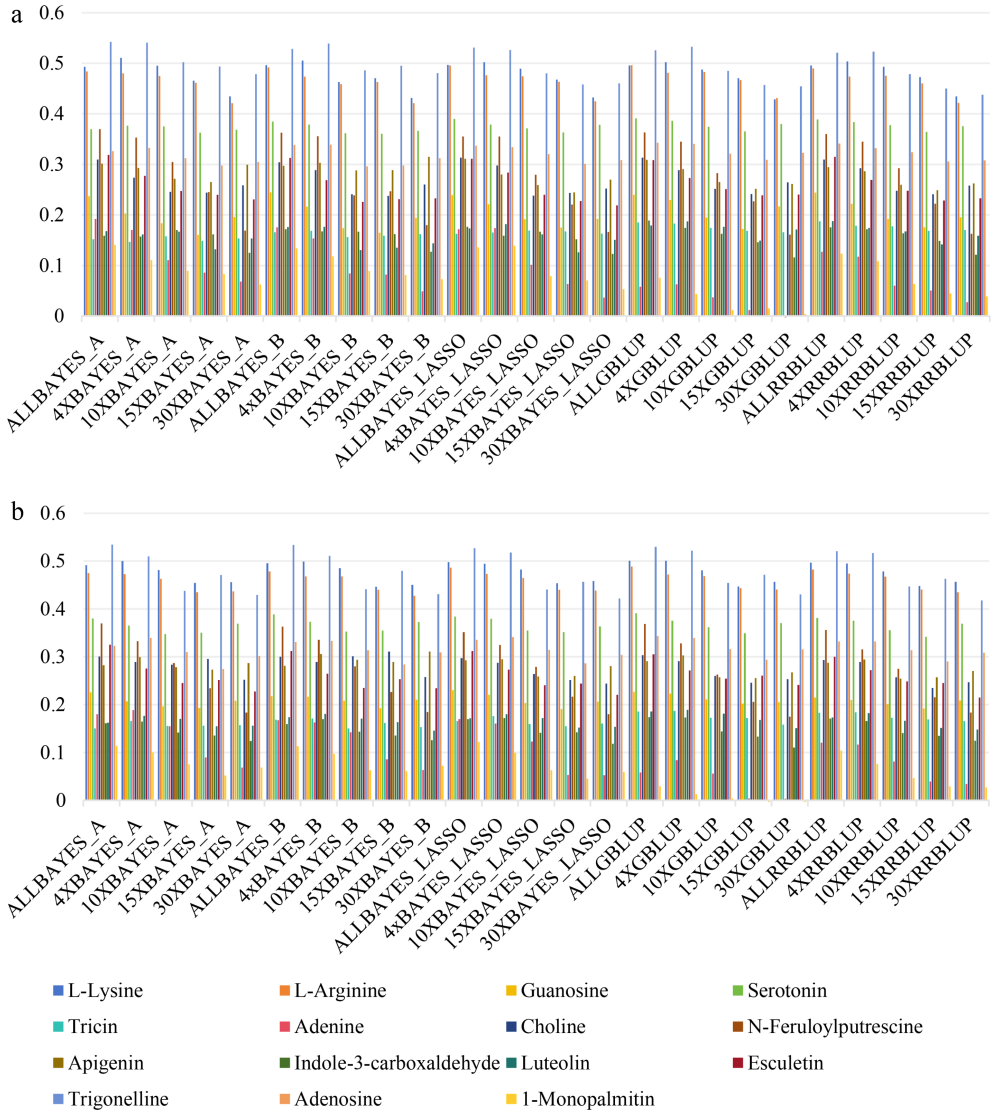


Fig. 3 The prediction accuracy of each metabolism in 524 rice samples in different models. (a) The prediction accuracy of each metabolite was not interpolated for genotypes. (b) The prediction accuracy of each metabolite after interpolation of genotypes.

Although the Bayesian B model could partially mitigate this decline, it could not reverse the trend. In summary, the objective of this study is to explore and establish corresponding methods to evaluate the feasibility of applying metabolic data in genomic selection (GS) prediction. Although the sample size of the independent validation set is limited, it provides a model for subsequent research. In future studies, we can increase the number of varieties, focus on collecting varieties at different growth stages, and consider special varieties. Moreover, in current analytical research, environmental factors are often not adequately taken into account, with most studies simply using genomic information to predict phenotypes. Given that the environment has a significant impact on phenotypic expression, future research should incorporate environmental factors as fixed effects into the genomic selection (GS) breeding models. This approach can effectively enhance the model's generalization ability, enabling it to have higher application value in various practical breeding scenarios.

Conclusions

This study successfully integrated metabolomics data into the genomic selection (GS) framework and explored the methodological feasibility of utilizing metabolomic data in GS prediction. By analyzing publicly available multi-omics data from 524 rice accessions, we systematically compared the predictive performance of five GS models and validated the biological value of metabolomics data in genomic prediction using an independent validation set of 12 core varieties. This work lays a theoretical foundation for developing integrated multi-omics intelligence-driven breeding models, provides a technical pathway for transitioning from 'single agronomic trait prediction' to 'multi-omics collaborative prediction,' and holds significant practical implications for accelerating crop precision breeding.

Author contributions

The authors confirm their contributions to the paper as follows: study conception and design: Xia C, Luo X, Zou M, Fan D, Xia Z; data collection: Lu Y, Liu Q, Hassan HM, Vladislav I; analysis and interpretation of results: Xia C, Luo X; draft manuscript preparation: Xia C, Luo X, Lu Y. All authors reviewed the results and approved the final version of the manuscript.

Data availability

The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Acknowledgments

This research was supported by the Guangxi Ministry of Science and Technology (Grant No. GuikeAA23062015), Project of Sanya Yazhou Bay Science and Technology City (Grant No. SCKJ-JYRC-2022-57), Hainan Yazhou Bay Seed Lab (Grant No. B23YQ0002), and High-performance Computing Platform of YZBSTCACC.

Conflict of interest

The authors declare that they have no conflict of interest.

Dates

Received 16 December 2025; Revised 17 April 2026; Accepted 9 May 2026; Published online 31 July 2026

References

[1] Zhang C, Jiang S, Tian Y, Dong X, Xiao J, et al. 2023. Smart breeding driven by advances in sequencing technology. *Modern Agriculture* 1(1):43–56

[2] Wallace JG, Rodgers-Melnick E, Buckler ES. 2018. On the Road to Breeding 4.0: unraveling the good, the bad, and the boring of crop quantitative genomics. *Annual Review of Genetics* 52:421–444

[3] Heffner EL, Sorrells ME, Jannink JL. 2009. Genomic Selection for Crop Improvement. *Crop Science* 49:1–12

[4] Jannink JL, Lorenz AJ, Iwata H. 2010. Genomic selection in plant breeding: from theory to practice. *Briefings in Functional Genomics* 9(2):166–177

[5] Desta ZA, Ortiz R. 2014. Genomic selection: genome-wide prediction in plant improvement. *Trends in Plant Science* 19(9):592–601

[6] Hickey LT, N Hafeez A, Robinson H, Jackson SA, Leal-Bertioli SCM, et al. 2019. Breeding crops to feed 10 billion. *Nature biotechnology* 37(7):744–754

[7] Lou RN, Jacobs A, Wilder AP, Therkildsen NO. 2021. A beginner's guide to low-coverage whole genome sequencing for population genomics. *Molecular Ecology* 30(23):5966–5993

[8] Das S, Abecasis GR, Browning BL. 2018. Genotype Imputation from Large Reference Panels. *Annual Review of Genomics and Human Genetics* 19:73–96

[9] Davies RW, Kucka M, Su D, Shi S, Flanagan M, et al. 2021. Rapid genotype imputation from sequence with reference panels. *Nature Genetics* 53(7):1104–1111

[10] Zou M, Xia Z. 2022. Hyper-seq: A novel, effective and flexible marker-assisted selection and genotyping approach. *The Innovation* 3(4):100254

[11] Xiao N, Pan C, Li Y, Wu, Y, Cai, Y, et al. 2021. Genomic insight into balancing high yield, good quality, and blast resistance of japonica rice. *Genome Biology* 22(1):283

[12] Cui Y, Li R, Li G, Zhang F, Zhu T, et al. 2020. Hybrid breeding of rice via genomic selection. *Plant Biotechnology Journal* 18(1):57–67

[13] Xu S, Xu Y, Gong L, Zhang Q. 2016. Metabolomic prediction of yield in hybrid rice. *The Plant Journal* 88(2):219–227

[14] Chen W, Gao Y, Xie W, Gong L, Lu K, et al. 2014. Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nature Genetics* 46(7):714–721

[15] Wang W, Yu Z, Meng J, Zhou P, Luo T, et al. 2020. Rice phenolamindes reduce the survival of female adults of the white-backed planthopper Sogatella furcifera. *Scientific Reports* 10(1):5778

[16] Xia J, Yamaji N, Ma JF. 2014. An appropriate concentration of arginine is required for normal root growth in rice. *Plant Signaling & Behavior* 9(4):e28717

[17] Liu F, Fang S, Wang Q, Wang H, Niu J, et al. 2024. Effects of different concentrations of exogenous amino acids on the growth and related physiological indexes of rice seedlings. *Crop Magazine* 2024(2):71–79

[18] Meuwissen TH, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829

[19] Jarquín D, Crossa J, Lacaze X, Du Cheyron P, Daucourt J, et al. 2014. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics* 127:595–607

[20] Costa-Neto G, Fritsche-Neto R, Crossa J. 2021. Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. *Heredity* 126:92–106

[21] Pérez-Martín JE, Bonte D, Osorio S, Posé D. 2026. Metabolic plasticity and adaptive evolution in Fragaria vesca: bridging wild diversity to crop improvement. *Frontiers in Plant Science* 16:1729002

[22] Yun J, Burnett AC, Rogers A, Des Marais DL. 2025. Genotype by environment interactions in gene regulation underlie the response to soil drying in the model grass *Brachypodium distachyon*. *Molecular Biology and Evolution* 42(10):msaf218

[23] Xu Y, Yang W, Qiu J, Zhou K, Yu G, et al. 2024. Metabolic marker-assisted genomic prediction improves hybrid breeding. *Plant Communications* 6(1):101199

[24] Xu S, Zhu D, Zhang Q. 2014. Predicting hybrid performance in rice using genomic best linear unbiased prediction. *Proceedings of the National Academy of Sciences of the United States of America* 111(34):12456−12461

[25] Zhang X, Pérez-Rodríguez P, Burgueño J, Olsen M, Buckler E, et al. 2017. Rapid Cycling Genomic Selection in a Multiparental Tropical Maize Population. *G3* 7(7):2315−2326

[26] Rasheed A, Hao Y, Xia X, Khan A, Xu Y, et al. 2017. Crop breeding chips and genotyping platforms: progress, challenges, and perspectives. *Molecular Plant* 10(8):1047−1064

[27] Guo Z, Yang Q, Huang F, Zheng H, Sang Z, et al. 2021. Development of high-resolution multiple-SNP arrays for genetic analyses and molecular breeding through genotyping by target sequencing and liquid chip. *Plant Communications* 2(6):100230

[28] DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* 43(5):491−498

[29] Alex Buerkle C, Gompert Z. 2013. Population genomics based on low coverage sequencing: how low should we go? *Molecular Ecology* 22(11):3028−3035

[30] Wang Q, He M, Zhou Y, Xu R, Liang T, et al. 2025. Hyper-seq technology and genome-wide selection breeding of soybeans. *Agronomy* 15(2):264

[31] Lu Y, Xia C, Wang Z, Liu Q, Zhu M, et al. 2025. Assessment of the prediction accuracy of genomic selection for rice amylose content and gel consistency. *Agronomy* 15(2):336

[32] Wang Z, Xia C, Lu Y, Liu Q, Zou M, et al. 2024. Optimizing genomic selection methods to improve prediction accuracy of sugarcane single-stalk weight. *Agronomy* 14(12):2842

[33] Kim K, Nawade B, Nam J, Chu S, Ha J, et al. 2022. Development of an inclusive 580K SNP array and its application for genomic selection and genome-wide association studies in rice. *Frontiers in Plant Science* 13:1036177

[34] Tanaka R, Lui-King J, Mandaharisoa ST, Rakotondramanana M, Ranaivo HN, et al. 2024. Correction: From gene banks to farmer's fields: using genomic selection to identify donors for a breeding program in rice to close the yield gap on smallholder farms. *Theoretical and Applied Genetics* 137(6):124

[35] Mahantesh, Ganesamurthy K, Das S, Saraswathi R, Gopalakrishnan C, et al. 2022. Analysis of the efficiency of genomic selection models for predicting sheath blight resistance in rice (*Oryza sativa* L.). *International Journal of Bio-resource and Stress Management* 13(3):268−275

[36] Huang M, Balimponya EG, Mgonja EM, McHale LK, Luzi-Kihupi A, et al. 2019. Use of genomic selection in breeding rice (*Oryza sativa* L.) for resistance to rice blast (*Magnaporthe oryzae*). *Molecular Breeding* 39:114

[37] Yang C, Shen S, Zhou S, Li Y, Mao Y, et al. 2022. Rice metabolic regulatory network spanning the entire life cycle. *Molecular Plant* 15(2):258−275