TYPE Technology and Code
PUBLISHED 06 March 2023
DOI 10.3389/fpls.2023.1131493

IP4GS: Bringing genomic selection analysis to breeders

Tong Li[†], Shan Jiang[†], Ran Fu, Xiangfeng Wang, Qian Cheng* and Shuqin Jiang*

Frontiers Science Center for Molecular Design Breeding, College of Agriculture and Biotechnology, China Agricultural University, Beijing, China

OPEN ACCESS

EDITED BY
Feng Cheng,
Insititute of Vegetables and Flowers (CAAS),
China

REVIEWED BY
Darko Jevremović,
Fruit Research Institute, Serbia
Sean Robert Asselin,
Agriculture and Agri-Food Canada (AAFC),
Canada
Wentao Zhang,
National Research Council Canada, Canada

*CORRESPONDENCE
Qian Cheng
qchengray@cau.edu.cn
Shuqin Jiang
wanshi0066@126.com

[†]These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to
Plant Bioinformatics,
a section of the journal
Frontiers in Plant Science

RECEIVED 25 December 2022
ACCEPTED 20 February 2023
PUBLISHED 06 March 2023

CITATION
Li T, Jiang S, Fu R, Wang X, Cheng Q and Jiang S (2023) IP4GS: Bringing genomic selection analysis to breeders.
Front. Plant Sci. 14:1131493.
doi: 10.3389/fpls.2023.1131493

Genomic selection (GS), a strategy to use genotypes to predict phenotypes *via* statistical or machine learning models, has become a routine practice in plant breeding programs. GS can speed up the genetic gain by reducing phenotyping costs and/or shortening the breeding cycles. GS analysis is complicated involving data clean up and formatting, training and test population analysis, model selection and evaluation, and parameter optimization. In addition, GS analysis also requires some programming skills and knowledge of statistical modeling. Thus, we need a more practical GS tools for breeders. To alleviate this difficulty, we developed the web-based platform IP4GS (https://ngdc.cncb.ac.cn/ip4gs/), which offers a user-friendly interface to perform GS analysis simply through point-and-click actions. IP4GS currently includes seven commonly used models, eleven evaluation metrics, and visualization modules, offering great convenience for plant breeders with limited bioinformatics knowledge to apply GS analysis.

KEYWORDS

bioinformatics, genomic selection, genotype-to-phenotype prediction, web-based platform, R shiny

Introduction

Polygenic traits of plants, such as grain yield (GY), flowering time (FT), and plant height (PH), are usually controlled by many minor effect genes. In such cases, traditional marker-assisted selection (MAS), which relies on statistical power to identify makers/genes-traits associations, cannot effectively be applied to expedite trait improvement of polygenic traits. (Collard and Mackill, 2008; Xu and Crouch, 2008). Genomic selection (GS) predicting phenotypes from genome-wide molecular markers, may act as a complementary approach for the improvements of polygenic traits. (Desta and Ortiz, 2014; Crossa et al., 2017). GS models use genome-wide genetic markers to predict phenotypes, which can maximumly capture phenotypic variation contributed by multiple minor effect genes. (Cerrudo et al., 2018). To perform GS analysis, first, a GS model is built by genotypic and phenotypic data from a training population. The model is then employed to predict phenotypes of a candidate population based on their genotypic data. (Meuwissen et al., 2001; Desta and Ortiz, 2014). The advent of advanced next-generation sequence technologies, e.g. genotyping by targeted sequencing (GBTS) has significantly reduced the cost of genotyping. (Guo et al., 2019). Wide application of GS to more and more plant species has become feasible since genotyping cost is no longer a bottleneck (Bauck et al., 2018; Belamkar et al., 2018; Atanda et al., 2021; Wang et al., 2021).

Most popularly GS models are linear regression (or parametric) regression models. These models include the best linear unbiased prediction (BLUP) method represented by RRBLUP (ridge regression BLUP) and the Bayesian method represented by BayesA and BayesB (Meuwissen et al., 2001; Endelman, 2011). In recent years, machine learning (ML)-based methods have been introduced to build GS models, such as support vector machine (SVM), random forest (RF), deep learning, and light gradient boost (LGB) algorithms (Blondel et al., 2015; Qiu et al., 2016; Ma et al., 2018; Yan et al., 2021). ML methods have been proposed to be better than linear models to incorporate nonlinear relationships between genotypes and phenotypes. However, ML may require relatively larger data size to achieve better performance and outperform linear models on more complex datasets. (González-Recio and Forni, 2011; Yan et al., 2021; Yan and Wang, 2022). In the practice of plant breeding at the current stage, most of the time, the size of training datasets is not big enough and the property of datasets is still relatively simple, because the high cost of phenotyping must still be considered as an important factor. In this situation, BLUP and Bayesian methods outperform most ML methods and have thus gained popularity in plant breeding (Tanaka and Iwata, 2018; Yan et al., 2021).

Among all available GS methods, RRBLUP, a “simple” model that assumes all marker effects possessing the same variance, is the most widely used approach in GS application. It has therefore become the baseline prediction model for the evaluation performance of other GS models. However, the assumption of equal variance of marker effects may be not realistic. Bayesian methods which assumed prior distributions for the variance of marker effects, has become another popular GS approach due to its potential for better estimating of marker effects. Currently, the majority of GS tools are command-line interfaces, requiring advanced data management skills, good programing skills and knowledge of statistical modeling, which restrict the practical use by breeders in the seed industry. There has therefore been an urgent need to develop a web-based platform with a user-friendly interface for facilitating the use of GS strategies among the plant breeding community. To solve this problem, we developed IP4GS, a web-based interactive platform for genomic selection. IP4GS includes seven GS models and eleven evaluation metrics to help users select optimal models for GS analysis. It also includes bioinformatics pipelines for preprocessing of genotypic data and visualization modules for population analysis. The functionality of all modules may be invoked simply by point-and-click actions through a user-friendly interface developed using shiny.

Methods

Demo datasets for developing IP4GS

The public dataset from a population of 1,404 maize F1 hybrids, which generated by crossing 1,404 inbred lines with an elite tester line Zheng58, was used to develop the web-based platform of IP4GS (Liu et al., 2020; Xiao et al., 2021). The genotypic data of the 1,404 F_1 hybrids includes 4,903 SNPs selected from a total of 14.8 million SNPs; details of SNP selection were previously described by Cheng et al. (Cheng et al., 2021). Phenotypes include flowering time (FT), plant height (PH), and grain yield (GY) (Xiao et al., 2021). The demo datasets of genotypes and phenotypes are accessible at https://github.com/furan2019/IP4GSdata.

Implementation of a web-based IP4GS platform

Shiny is an R-based framework, allowing programmers to develop interactive web applications (https://www.rstudio.com/products/shiny/). Benefitting from the expandability and useability, Shiny has been widely used to develop online applications for bioinformatics software or interactive plot tools (McMurdie and Holmes, 2015; Yu et al., 2018; Sievert, 2020). The shiny-based application can be accessed on the local host or deployed on the public internet for public access. A shiny-based application usually consists of two parts, a user-interface (UI) script (IP4GS_UI.r for IP4GS) and a server script (IP4GS_server.r for IP4GS). The UI script controls the layout of different panels and visualization of results and bridges the user inputs and background functions. IP4GS utilized several packages to enrich and improve the UI interactive experience, such as “DT” for dynamic tables, “plotly” for dynamic plots, “shinycssloaders” for loading animations, “shinybusy” for progress notification, and “shinyWidgets” for input of multiform parameters and dynamic controls. HTML5 language and condition panels were also introduced to improve and optimize the layout of panels. The server script plays an important role in shiny-based applications. All functions provided by IP4GS were achieved by server script, including data input, data preprocessing, and GS model building and evaluation. For real-time interaction, user-defined parameters and operations are passed to the server script, which then executes the corresponding functions and formats the outputs. Lastly, the server script returns the results to a specific location according to flags that can bridge the UI script and server script.

GS models and evaluation metrics in the GS analysis panel

IP4GS integrates five linear models and two ML methods for GS application, which are all called from R libraries (Table 1). The RRBLUP method is called from the package “rrBLUP”. The RRBLUP model is based on the assumption that all markers have equal variance with small and nonzero effect (Endelman, 2011). The three Bayesian methods, BayesA, BayesB, and BayesC, are called from the package “BGLR (Bayesian generalized linear regression)” (de los Campos and Pérez, 2015). BayesA utilizes a scaled *t*-distribution for estimating marker effects. BayesB is similar to BayesA, with the main difference that it utilizes both shrinkage and variable selection algorithms to estimate marker effects. By

TABLE 1 Seven GS methods integrated into IP4GS.

Model	Important parameter (s)	Package	Reference
BayesA	nlter burnIn thin	BGLR	(de los Campos and Pérez, 2015)
BayesB			
BayesC			
Least absolute shrinkage and selection operator (LASSO)	alpha	glmnet	(Friedman et al., 2010)
Ridge regression best linear unbiased prediction (RRBLUP)	N/A	rrBLUP	(Endelman, 2011)
Support vector regression (SVR)	gamma; cost; kernel	e1071	(Meyer et al., 2014)
Random forest regression (RFR)	ntree; nodesize	randomForest	(Liaw and Wiener, 2002)

contrast, BayesC estimates marker effects based on a Gaussian distribution. The LASSO (least absolute shrinkage and selection operator) method is called from the package "glmnet (Lasso and elastic-net regularized generalized linear models)," which combines both shrinkage and variable selection algorithms to estimate marker effects. The two ML methods, SVR (support vector regression) and RFR (random forest regression), are called from the "e1071" and "randomForest" packages, respectively. SVR finds an appropriate line (or hyperplane in higher dimensions) to fit the data, and RFR uses a regression model rooted in bootstrapping sample observations (Liaw and Wiener, 2002; Meyer et al., 2014; Ornella et al., 2014).

To enable comprehensive evaluation of the prediction accuracy of selected GS models, IP4GS integrates eleven evaluation metrics: five correlation-based metrics to globally measure the relationship between observed and predicted phenotypes, and six threshold-based metrics to count accurately predicted top-ranked individuals (Table 2). The correlation-based methods are Pearson correlation coefficient (PCC), Kendall rank correlation coefficient (KCC), Spearman rank correlation coefficient (SCC), squared R coefficient of determination (R^2), and mean squared error (MSE). These metrics usually measure the global performance of models. For example, KCC treats all pairs equally; however, in breeding practice, more attention should be paid to extreme values such as high yield and short flowering time (Blondel et al., 2015). Thus, six threshold-based metrics were introduced for top-k individuals with ideal phenotypic value. These are normalized discounted cumulative gain (NDCG), mean NDCG, relative efficiency (RE), Accuracy, F-score, and Cohen's kappa coefficient (Kappa) (Ornella et al., 2014; Blondel et al., 2015). The calculation methods for these metrics are described in Table 2. X is an array of observed phenotypic values, and Y is an array of predicted phenotypic values. For Accuracy, F-score, and Kappa, positive samples are those with ideal phenotypic value.

Bioinformatics pipelines in the data preprocessing panel

Currently, IP4GS accepted genotypes with AA/AB/BB alleles (allele format) and 0/1/2 format (numeric format). When data processing button is pressed, the data processing function will compute allele frequency, define the major and minor alleles, filter and format submitted genotypic data. It is worth noting that genotypic data in numeric format is not applicable to this function, IP4GS suppose that genotypic data submitted in numeric format is already processed by users. For genotypic data in

allele format, the function will compute allele frequency and missing rate, define the major and minor alleles. Genotypic data will be filtered by minor allele frequency (MAF) and missing rate, markers with MAF below 0.05 (<= 0.05) or missing rate higher than 0.2 (>= 0.2) will be removed and the criteria can be defined by users. For format conversion, IP4GS uses a common 0, 1, 2 coding scheme based on defined major and minor alleles, AA (homozygous genotype comprising major alleles), AB (homozygous genotype) and BB (homozygous genotype comprising minor alleles) will be coded as 0, 1, and 2, respectively. Two methods are provided in IP4GS for imputation of missing genotype values, "Mean" for mean value of each SNP and "Major" for the code with highest frequency of each SNP. In addition, IP4GS can implement dimensionality reduction of genotypic data using three algorithms, the "prcomp" function with default parameters (e.g., center = TRUE and scale. = FLASE) in the "stats" package for principal component analysis (PCA), the "umap" package for uniform manifold approximation and projection (UMAP), and the "tsne" package for t-distributed stochastic neighbor embedding (t-SNE). For current version of IP4GS, the utilization of all above three algorithms with default parameters except the dimension which can be defined by users.

Permissions and accessibility

The free version of IP4GS for academia distributed in the public domain is available at https://ngdc.cncb.ac.cn/ip4gs/. All functions of the public IP4GS described herein are freely accessible for small datasets. There are suggested limitations on the number of SNP markers (<10,000 SNPs) and samples (<1,000 individuals) because of the limited high-performance computing resources of the public web server hosting IP4GS. Considering the need for confidentiality of breeding data by industrial users, an offline version of IP4GS that can be installed on a private server or local devices without any limitation on the number of SNPs and samples is also available. Users interested in nonlimited IP4GS may contact the corresponding author for access.

TABLE 2 Eleven metrics for model evaluation integrated into IP4GS.

Evaluation Metric	Formula	Remarks
Pearson correlation coefficient (PCC, r, R)	$\mathrm{PCC}(X,Y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$	$\bar{x}$ is the mean of X, $\bar{y}$ is the mean of Y
Kendall rank correlation coefficient (KCC, tau, τ)	$\mathrm{KCC}\ (X,Y)=\frac{N1-N2}{n(n-1)/2}$	$N1$: number of concordant pairs (e.g., x_i and y_i); $N2$: number of discordant pairs
Spearman rank correlation coefficient (SCC, rho, ρ)	$\mathrm{SCC}(X,Y)=\frac{\mathrm{cov}(R(X),R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}}$	$R(X)$ and $R(Y)$ are the rank of variables
Coefficient of determination, R squared (R^2, r^2)	$R2\ (X,\ Y)=\ 1-\frac{\sum_{i=1}^{n}(x_i-y_i)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$	$\bar{x}$ is the mean of X
Mean squared error (MSE)	$\mathrm{MSE}(X,Y)\ =\frac{\sum_{i=1}^{n}(x_i-y_i)^2}{N}$	N is the total number of samples
Normalized discounted cumulative gain (NDCG)	$\mathrm{NDCG}(k,\ X,\ Y)=\frac{\sum_{i=1}^{k}x(i,Y)d(i)}{\sum_{i=1}^{k}x(i,X)d(i)}$	$d(i)=\frac{1}{\log_2 i+1}$is a discount function; $x(i,Y)$ is the ith value of X with the order of Y; $x(i,X)$ is the ith value of X with the order of X; k is the top k individuals with ideal phenotypic values; $\bar{x}$ is the mean of X
Mean normalized discounted cumulative gain (mean NDCG)	$\mathrm{meanNDCG@}K(X,Y)=\frac{1}{K}\sum_{k=1}^{K}\mathrm{NDCG}(k,X,Y)$	
Relative efficiency (RE)	$\mathrm{RE}(k)=\frac{\frac{1}{k}\sum_{i=1}^{k}x(i,Y)-\bar{x}}{\frac{1}{k}\sum_{i=1}^{k}x(i,X)-\bar{x}}$	
Accuracy	$\mathrm{Accuracy}=\frac{N_{\mathrm{TP}}+N_{\mathrm{TN}}}{N}$	TP: true positive; TN: true negative; FP: false positive; FN: false negative; default $\beta=1$; $P=\frac{N_{\mathrm{TP}}}{N_{Op}}$; $R=\frac{N_{\mathrm{TP}}}{N_{Pp}}$; $N_{Op}=N_{\mathrm{TP}}+N_{\mathrm{FN}}$; $N_{On}=N_{\mathrm{FP}}+N_{\mathrm{TN}}$; N_{Pp} $=N_{\mathrm{TP}}+N_{\mathrm{FN}}$; $N_{Pn}=N_{\mathrm{FN}}+N_{\mathrm{TN}}$.
F-score	$F_{\beta}-\mathrm{score}=\frac{(1+\beta^2)\times P\times R}{\beta^2\times P+\ R}$	
Cohen's kappa coefficient (Kappa)	$\mathrm{Kappa}=\frac{p_0-p_e}{1-p_e}$; $p_0=\frac{\mathrm{N_{TP}}+\ \mathrm{N_{TN}}}{N}$; $p_e=\frac{N_{Pp}}{N}\times\frac{N_{Op}}{N}+\frac{N_{Pn}}{N}\times\frac{N_{On}}{N}$	

Results

Overall workflow

The IP4GS platform can be divided into two main panels of functional modules: the "Data preprocessing" panel and the "GS analysis" panel. The Data preprocessing panel comprises not only bioinformatics pipelines for data processing and quality control of input datasets, but also a variety of dimensionality reduction (DR) algorithms for population structure visualization based on genotypes (Figure 1, left). Additionally, users may either preview the processed data through the web browser or download the data to a local computer. The GS analysis panel takes the input datasets generated from the first panel and performs G2P prediction with seven regression-based GS models and eleven evaluation metrics (Figure 1, right). The parameters for each GS model can be predefined on the parameter input panel by users. Visualization of evaluation results allows users to identify and output predicted phenotypes from the optimal model.

Functional modules in the data preprocessing panel

To run the analytical modules in the Data preprocessing panel, users need to prepare one mandatory input file of genotypic and phenotypic data and one optional file containing data for fixed effects considered by the model (Figure 2A). Acceptable data formats for the genotypic data file are either the standard "HapMap" format or a plain text file containing a matrix of genotypes in columns and individuals in rows. In the matrix file, the genotype of each SNP must be converted by the users to "0," "1," or "2," representing homozygous major alleles, heterozygous alleles, and homozygous minor alleles, respectively. If users select the "Custom" option from the pulldown list of file formats, the genotype of each SNP can be entered in the character format "A, C, G, T," which is automatically converted to the "0, 1, 2" format on the basis of allele frequency computed by IP4GS. Either a.csv (comma-separated values) file or a tab-delimited.txt (text) file is acceptable by IP4GS for phenotypic data.

After the input files are uploaded, the genotypic data file is first processed with regard to two criteria, namely minor allele frequency (MAF, default<= 0.05) and missing rate (MR, default >= 0.2), to remove low-quality SNPs (Figure 2B). In addition, IP4GS offers users the option of whether imputation is performed on genotypic data or not. As reference haplotype-based imputation may consume a large volume of computing resources, IP4GS only offers a simplified imputation method using "Mean" or "Major" to replace missing genotypic values (see Methods). The preprocessed and filtered genotypic data can then be partially previewed or fully downloaded by clicking the download button in the console (Figure 2C). In addition to the raw and processed genotypic and phenotypic data, the DR data computed using PCA (principal component analysis), UMAP

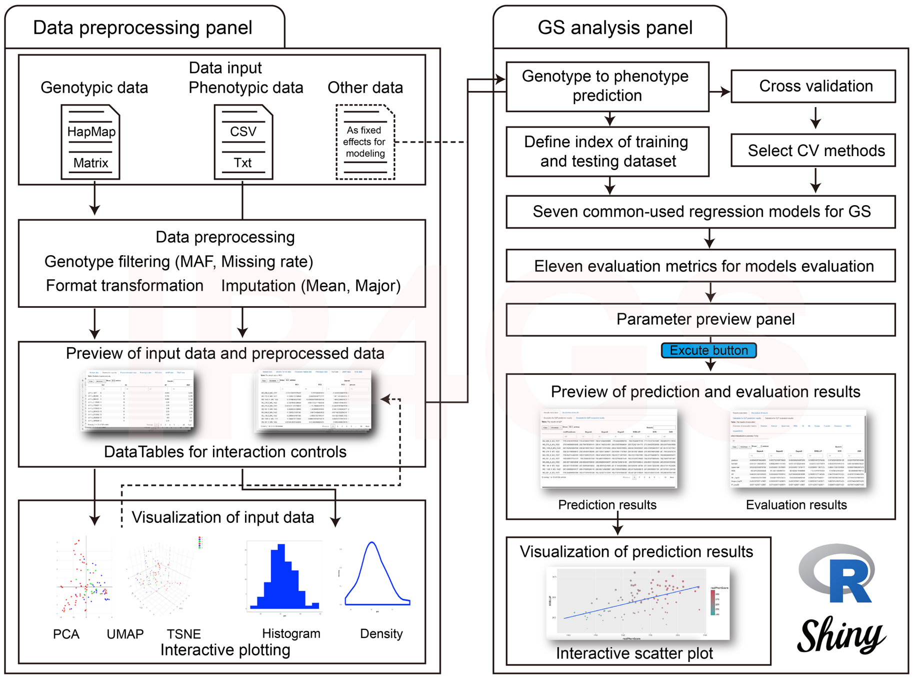
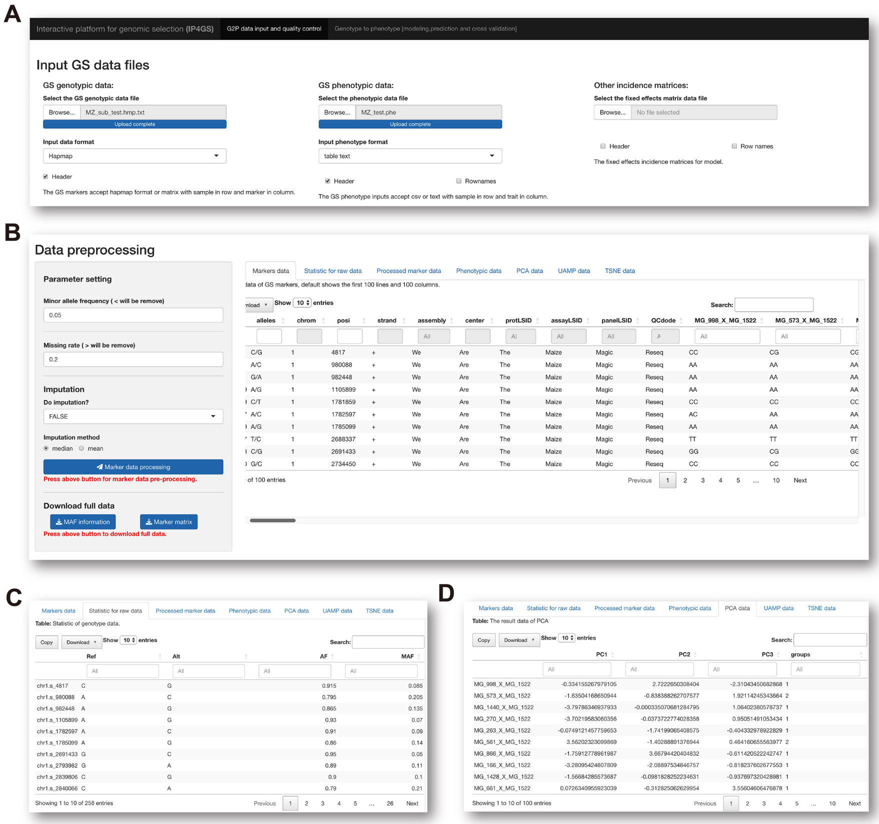

Data preprocessing panel
Data input
Genotypic data
Phenotypic data
Other data
HapMap
Matrix
CSV
Txt
As fixed effects for modeling
Data preprocessing
Genotype filtering (MAF, Missing rate)
Format transformation
Imputation (Mean, Major)
Preview of input data and preprocessed data
DataTables for interaction controls
Visualization of input data
PCA
UMAP
TSNE
Histogram
Density
Interactive plotting
GS analysis panel
Genotype to phenotype prediction
Cross validation
Define index of training and testing dataset
Select CV methods
Seven common-used regression models for GS
Eleven evaluation metrics for models evaluation
Parameter preview panel
Excute button
Preview of prediction and evaluation results
Prediction results
Evaluation results
Visualization of prediction results
Interactive scatter plot
R
Shiny

FIGURE 1
Overall workflow of IP4GS. IP4GS comprises two panels of functional modules: the Data preprocessing panel (left) and the GS analysis panel (right).



Input GS data files

GS genotypic data:
Select the GS genotypic data file
MZ_sub_test.hmp.txt
Input data format
Hapmap
☑ Header
The GS markers accept hapmap format or matrix with sample in row and marker in column.

GS phenotypic data:
Select the phenotypic data file
MZ_test.phe
Input phenotype format
table text
☑ Header ☐ Rownames
The GS phenotype inputs accept csv or text with sample in row and trait in column.

Other incidence matrices:
Select the fixed effects matrix data file
No file selected
☐ Header ☐ Row names
The fixed effects incidence matrices for model.

Data preprocessing

Parameter setting
Minor allele frequency (< will be remove)
0.05
Missing rate (> will be remove)
0.2
Imputation
Do imputation?
FALSE
Imputation method
● median ○ mean
Press above button for marker data pre-processing.
Download full data
Press above button to download full data.

ɟata of GS markers, default shows the first 100 lines and 100 columns.

Show 10 entries

alleles	chrom	posi	strand	assembly	center	protLSID	assayLSID	panelLSID	QCdode	MG_998_X_MG_1522	MG_573_X_MG_1522
C/G	1	4817	+	We	Are	The	Maize	Magic	Reseq	CC	CG
A/C	1	980088	+	We	Are	The	Maize	Magic	Reseq	AA	AA
G/A	1	982448	+	We	Are	The	Maize	Magic	Reseq	AA	AA
A/G	1	1105899	+	We	Are	The	Maize	Magic	Reseq	AA	AA
C/T	1	1781859	+	We	Are	The	Maize	Magic	Reseq	CC	CC
A/C	1	1782597	+	We	Are	The	Maize	Magic	Reseq	AC	AA
A/G	1	1785099	+	We	Are	The	Maize	Magic	Reseq	AA	AA
T/C	1	2688337	+	We	Are	The	Maize	Magic	Reseq	TT	TT
C/G	1	2691433	+	We	Are	The	Maize	Magic	Reseq	GG	CG
G/C	1	2734450	+	We	Are	The	Maize	Magic	Reseq	CC	CC

of 100 entries

Table: Statistic of genotype data.

Show 10 entries

	Ref	Alt	AF	MAF
chr1.s_4817	C	G	0.915	0.085
chr1.s_980088	A	C	0.795	0.205
chr1.s_982448	A	G	0.865	0.135
chr1.s_1105899	A	G	0.93	0.07
chr1.s_1782597	A	C	0.91	0.09
chr1.s_1785099	A	G	0.86	0.14
chr1.s_2691433	G	C	0.95	0.05
chr1.s_2793982	G	A	0.89	0.11
chr1.s_2839806	C	G	0.9	0.1
chr1.s_2840066	C	A	0.79	0.21

Showing 1 to 10 of 258 entries

Table: The result data of PCA

Show 10 entries

	PC1	PC2	PC3	groups
MG_998_X_MG_1522	-0.334155267979105	2.7222650308404	-2.31043450682868	1
MG_573_X_MG_1522	-1.63504168650944	-0.838388262707577	1.92114245343664	2
MG_1440_X_MG_1522	-3.79786346937933	-0.000335070681284795	1.06402380578737	1
MG_270_X_MG_1522	-3.70219583060356	-0.0373722774028358	0.95051491053434	1
MG_263_X_MG_1522	-0.0749121457759653	-1.74199065408575	-0.404332978922829	1
MG_561_X_MG_1522	3.56202323099869	-1.40288891376944	0.464160655563977	2
MG_866_X_MG_1522	-1.75912778961987	3.66794420404832	-0.611420522242747	1
MG_166_X_MG_1522	-3.28095424807809	-2.08897534846757	-0.818237602677553	1
MG_1428_X_MG_1522	-1.56684285573687	-0.0981828252234631	-0.937697320428981	1
MG_661_X_MG_1522	0.0726340955923039	-0.312825062629954	3.55604606476878	1

Showing 1 to 10 of 100 entries

FIGURE 2
Preprocessing of genotypic data. **(A)** IP4GS accepts genotypic data in HapMap, Matrix, and Custom file formats, phenotypic data in CSV and tab-delimited text formats, and an optional file including features as fixed effects. **(B)** Control console for genotypic data filtration (left) and display window for preview of a variety of data tables (right). **(C)** Preview of processed genotypic data with statistics for MAF, AF, and MR for each SNP. **(D)** Preview of DR data generated by PCA, UMAP, and t-SNE algorithms.

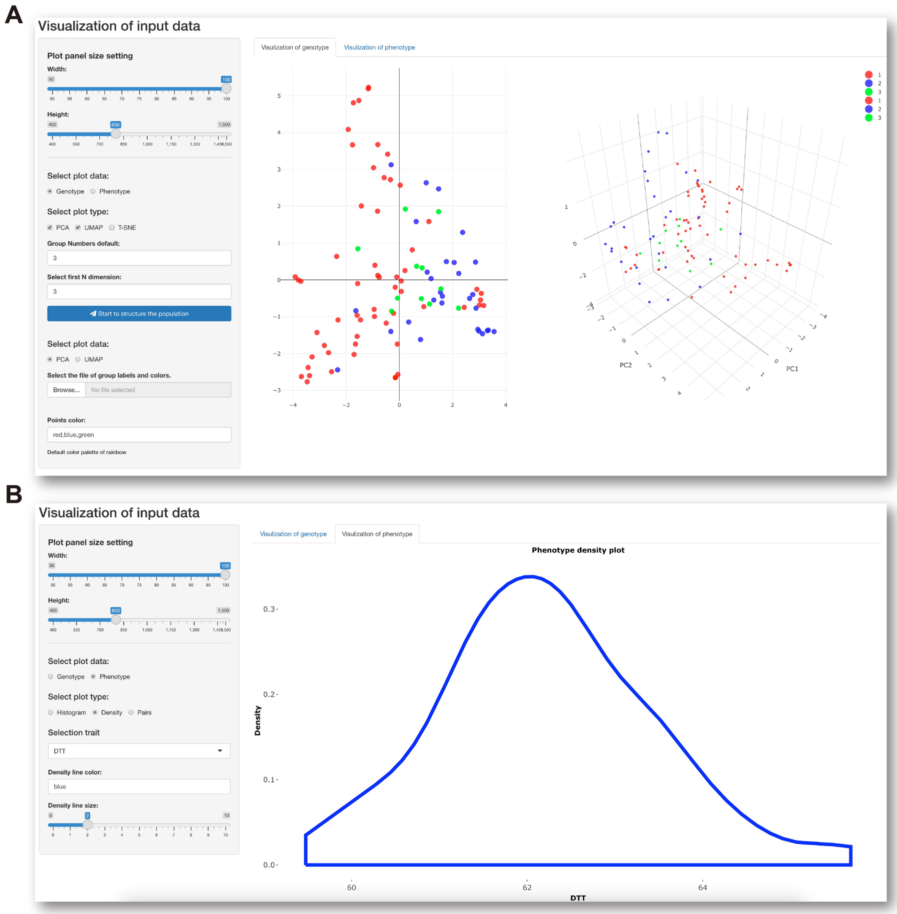
(uniform manifold approximation and projection), and *t*-SNE (*t*-distributed stochastic neighbor embedding) algorithms can also be previewed and downloaded (Figure 2D).

The last module in the Data preprocessing panel is interactive visualization of genotypic and phenotypic data with a control console and display window. This function facilitates not only visualization of the genetic composition of the population subjected to GS analysis (Figure 3A) but also understanding of the distribution of phenotypes about to be predicted (Figure 3B). When all data preprocessing is complete, IP4GS performs a quality-control assessment on the processed data to ensure proper execution of the subsequent GS analysis.

Functional modules in the GS analysis panel

After the completion of quality control, users may proceed to the GS analysis panel. The current version of IP4GS supports seven GS methods commonly used in plant breeding: five linear methods (RRBLUP, BayesA, BayesB, BayesC, and LASSO) and two ML methods (the RFR and SVR algorithms) (Table 1). We highly recommend that users try all seven methods for initial evaluation of G2P prediction results for a given set of data since it has previously been reported that no single GS method is superior for all traits and species (Ornella et al., 2014; Yan et al., 2021; Robert et al., 2022). It is important to select the optimal model given a designated trait and species to ensure the most precise prediction. Additionally, IP4GS offers eleven evaluation metrics for comprehensive evaluation of model performance; these include not only correlation-based Pearson, Spearman, and Kendall algorithms but also other algorithms such as *F*-score and MSE (Table 2).

The GS analysis panel is composed of five major parts: modeling console, parameter display window, results display window, visualization console, and plot display window. Modeling console consists of G2P console and CV console. From the G2P console, users may select GS methods and corresponding arguments, define indexes of training and test samples, and select evaluation metrics and corresponding arguments (Figure 4A). The CV console offers the option of three commonly used CV methods: *k*-fold, holdout, and leave-one-out schemes. Users may also set up the repeat time for CV and proportion of testing set included from the console. As long as all parameters for G2P models and CV methods are set, the display window will exhibit these preset parameters for users to



A

Visualization of input data

Plot panel size setting

Width: 100

Height: 800

Select plot data:

● Genotype ○ Phenotype

Select plot type:

☑ PCA ☑ UMAP ☐ T-SNE

Group Numbers default:

3

Select first N dimension:

3

Start to structure the population

Select plot data:

● PCA ○ UMAP

Select the file of group labels and colors.

Browse... No file selected

Points color:

red,blue,green

Default color palette of rainbow

B

Visualization of input data

Plot panel size setting

Width: 100

Height: 800

Select plot data:

○ Genotype ● Phenotype

Select plot type:

○ Histogram ● Density ○ Pairs

Selection trait

DTT

Density line color:

blue

Density line size: 2

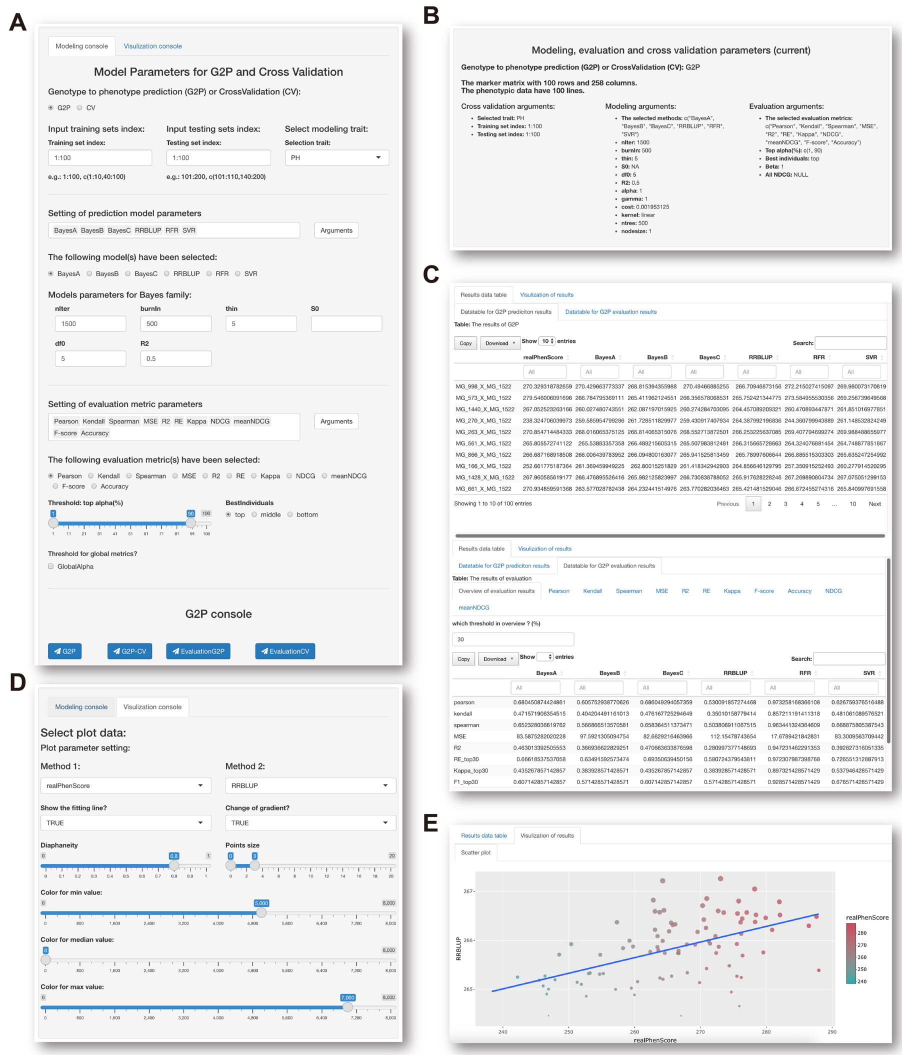
FIGURE 3
Data visualization console. **(A)** Parameter-setting console (left) for visualization of population structure (right) based on genotypic data. **(B)** Parameter-setting console (left) for visualization of data distribution (right) of selected phenotypic data.

double check (Figure 4B). If no further corrections are needed, users may press the execution button to run the GS analysis. It is worth noting that users may select all seven GS methods and eleven evaluation metrics and run G2P prediction and CV evaluation simultaneously to generate a table containing prediction and evaluation results for comparison.

From the results display window, G2P prediction and model evaluation results derived from all GS methods and evaluation metrics selected can be previewed before downloading of the full results (Figure 4C). Users may further use the visualization console to visually compare either observed phenotypes and predicted phenotypes or any two sets of predicted phenotypes from any two selected methods (Figure 4D). When parameters are set up on the visualization console, a scatter plot depicting the correlation of observed and predicted phenotypes is generated on the plot display window (Figure 4E). When IP4GS finishes the analysis of a set of breeding data, users may select the best prediction results from the optimal model to download.

Discussion

Owing to the rapid advancement of next-generation sequencing, GBTS has greatly reduced the expense of genotyping, making GS-assisted breeding more and more feasible for a growing number of plant species. However, GS analysis requires not only basic bioinformatics skills for data management but also experience in data modeling. The IP4GS platform was developed using the R shiny package, as an interactive, user-friendly web interface, allowing breeders perform GS analysis without the need of bioinformatics skills. However, as with any web-based application, limitations exist. We only integrated seven GS methods commonly used in plant breeding into



Model Parameters for G2P and Cross Validation

Genotype to phenotype prediction (G2P) or CrossValidation (CV):

● G2P ○ CV

Input training sets index: Training set index: 1:100 (e.g.: 1:100, c(1:10,40:100))

Input testing sets index: Testing set index: 1:100 (e.g.: 101:200, c(101:110,140:200))

Select modeling trait: Selection trait: PH

Setting of prediction model parameters: BayesA BayesB BayesC RRBLUP RFR SVR

The following model(s) have been selected:

● BayesA ○ BayesB ○ BayesC ○ RRBLUP ○ RFR ○ SVR

Models parameters for Bayes family:

nIter: 1500 burnIn: 500 thin: 5 S0: df0: 5 R2: 0.5

Setting of evaluation metric parameters: Pearson Kendall Spearman MSE R2 RE Kappa NDCG meanNDCG F-score Accuracy

The following evaluation metric(s) have been selected:

● Pearson ○ Kendall ○ Spearman ○ MSE ○ R2 ○ RE ○ Kappa ○ NDCG ○ meanNDCG ○ F-score ○ Accuracy

Threshold: top alpha(%): 1 – 90 (1 to 100)

BestIndividuals: ● top ○ middle ○ bottom

Threshold for global metrics?

□ GlobalAlpha

G2P console

Modeling, evaluation and cross validation parameters (current)

Genotype to phenotype prediction (G2P) or CrossValidation (CV): G2P

The marker matrix with 100 rows and 258 columns.
The phenotypic data have 100 lines.

Cross validation arguments:

- **Selected trait:** PH
- **Training set index:** 1:100
- **Testing set index:** 1:100

Modeling arguments:

- **The selected methods:** c("BayesA", "BayesB", "BayesC", "RRBLUP", "RFR", "SVR")
- **nIter:** 1500
- **burnIn:** 500
- **thin:** 5
- **S0:** NA
- **df0:** 5
- **R2:** 0.5
- **alpha:** 1
- **gamma:** 1
- **cost:** 0.001953125
- **kernel:** linear
- **ntree:** 500
- **nodesize:** 1

Evaluation arguments:

- **The selected evaluation metrics:** c("Pearson", "Kendall", "Spearman", "MSE", "R2", "RE", "Kappa", "NDCG", "meanNDCG", "F-score", "Accuracy")
- **Top alpha(%):** c(1, 90)
- **Best individuals:** top
- **Beta:** 1
- **All NDCG:** NULL

Table: The results of G2P

Show 10 entries Search:

	realPhenScore	BayesA	BayesB	BayesC	RRBLUP	RFR	SVR
	All	All	All	All	All	All	All
MG_998_X_MG_1522	270.329318782659	270.429663773337	268.815394355988	270.49466885255	266.70946873156	272.215027415097	269.980073170819
MG_573_X_MG_1522	279.546006091696	266.784795369111	265.411962124551	266.356578068531	265.752421344775	273.584955530356	269.256739649568
MG_1440_X_MG_1522	267.052523263166	260.027480743551	262.087197015925	260.274284703095	264.457089209321	260.470893447871	261.851016977851
MG_270_X_MG_1522	238.324706039973	259.585954799286	261.728511829977	259.430917407934	264.387992196836	244.360799943889	261.148532824249
MG_263_X_MG_1522	270.854714484333	268.016065375125	266.814065315076	268.552713872501	266.253225637085	269.407794699274	269.988488655977
MG_561_X_MG_1522	265.805572741122	265.53883357358	266.489215605315	265.507983812481	266.315665728663	264.324076881454	264.748877851867
MG_866_X_MG_1522	266.687168918508	266.006439783952	266.094800163077	265.941525813459	265.78997606644	266.885515303303	265.635247254992
MG_166_X_MG_1522	252.661775187364	261.369459949225	262.80015251829	261.418342942903	264.856646129795	257.350915252493	260.277914520295
MG_1428_X_MG_1522	267.960585619177	266.476895526416	265.982125823997	266.730838788052	265.917628228246	267.269890804734	267.075051299153
MG_661_X_MG_1522	270.934859591368	263.577028782438	264.232441514976	263.770282036463	265.421481529046	266.672455274316	265.840997691558

Showing 1 to 10 of 100 entries

Table: The results of evaluation

Overview of evaluation results | Pearson | Kendall | Spearman | MSE | R2 | RE | Kappa | F-score | Accuracy | NDCG | meanNDCG

which threshold in overview ? (%)

30

Show entries Search:

	BayesA	BayesB	BayesC	RRBLUP	RFR	SVR
	All	All	All	All	All	All
pearson	0.680450874424861	0.605752938770626	0.686049294057359	0.530091857274468	0.973258168366108	0.626759376516488
kendall	0.471571906354515	0.404204491161013	0.476167725294649	0.35010158779414	0.857211191411318	0.481061089576521
spearman	0.652328036619762	0.566866513570581	0.658364511373471	0.503806911067515	0.963441324364609	0.668875805387543
MSE	83.5875282020228	97.5921305094754	82.6629216463966	112.15478743654	17.6789421842831	83.3009563709442
R2	0.463013392505553	0.366936622829251	0.470663633876598	0.280997377148693	0.947231462291353	0.392827316051335
RE_top30	0.66618537537058	0.63491592573474	0.69350639450156	0.580724379543811	0.972307987398768	0.726551312887913
Kappa_top30	0.435267857142857	0.383928571428571	0.435267857142857	0.383928571428571	0.897321428571429	0.537946428571429
F1_top30	0.607142857142857	0.571428571428571	0.607142857142857	0.571428571428571	0.928571428571429	0.678571428571429

Select plot data:

Plot parameter setting:

Method 1: realPhenScore

Method 2: RRBLUP

Show the fitting line? TRUE

Change of gradient? TRUE

Diaphaneity: 0.8 (0 to 1)

Points size: 0 – 3 (0 to 20)

Color for min value: 5,000 (0 to 8,000)

Color for median value: 0 (0 to 8,000)

Color for max value: 7,000 (0 to 8,000)

FIGURE 4
GS analysis panel. **(A)** Modeling console for selection of GS methods, evaluation metrics, and setting of model parameters. **(B)** Display window for viewing selected models and parameter settings. **(C)** Previews of prediction results from multiple GS methods (upper panel) and evaluation results from different metrics (bottom panel). **(D)** Visualization console for viewing prediction results. **(E)** Display window for viewing scatter plots of observed and predicted phenotypes.

IP4GS. It is impossible to include all existing methods, especially those ML methods that require intensive computing resources for model training and parameter tuning. The seven methods were selected on the basis of a previously published evaluation of multiple statistical and ML methods, and all seven satisfy three basic criteria (Yan et al., 2021). First, prediction accuracy may not be greatly reduced when the size of the training set is smaller than that of the testing set, since the ratio of training versus test set is usually 1:4 in the seed industry (Yan et al., 2021). Second, model training and CV evaluation may not require too much CPU and memory usage. Third, non-excessive parameters and manual model-tuning are required to properly perform the GS analysis. Another common issue for all web-based applications is the upper size limit of input files uploaded for GS analysis. It is better for users to compile marker sets containing less than 10,000 SNPs, and a population size of smaller than 1,000. Therefore, if a user wants to perform GS analysis of a large dataset or use ML methods consuming intensive computing resource, we do not recommend using IP4GS. Furthermore, IP4GS, as a GS analysis platform, is theoretically applicable to other crops that have successfully applied the GS strategy including rice and wheat. And other species which can provide same format of genotypic and phenotypic data are also applicable but the effectiveness needs further investigation and exploration.

Previously reports indicate that no single GS method outperformed others for all evaluated traits and species. (Heffner et al., 2010; Xiao et al., 2021; Yan et al., 2021). The only solution is to evaluate multiple GS methods and select the optimal one for specific traits and species. Given this need, we integrated multiple GS methods and evaluation metrics so that users may compare results from different predictive models. The seven methods we selected usually generate similar prediction results according to our previous evaluation. Hence, the current version of IP4GS does not include a solution for integration of multi-model prediction results. However, in case of multi-model prediction, we also provided an R script to integrate prediction results from two algorithms. The tool is freely available at https://github.com/furan2019/IP4GSdata.git for users to integrate multi-model prediction results.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding authors.

Author contributions

SQJ and QC conceived and supervised the project. QC, SQJ, and XFW wrote the manuscript. TL developed the main platform and genomic selection modules of IP4GS. SJ developed the bioinformatics pipeline and analytical modules. RF developed the evaluation modules and visualization modules. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Hainan Yazhou Bay Seed Laboratory (B21HJ0505), the Chinese Universities Scientific Fund (2022TC139), and the 2115 Talent Development Program of China Agricultural University.

Acknowledgments

We appreciate the China National Center for Bioinformation at the Beijing Institute of Genomics for offering public server domain and high-performance computing resources for hosting the IP4GS platform. We also appreciate Ms. Yang Zhang for distributing the codes of IP4GS on the server.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Atanda, S. A., Olsen, M., Burgueño, J., Crossa, J., Dzidzienyo, D., Beyene, Y., et al. (2021). Maximizing efficiency of genomic selection in CIMMYT's tropical maize breeding program. *Theor. Appl. Genet.* 134, 279–294. doi: 10.1007/s00122-020-03696-9

Bauck, S., Lee, J., Morota, G., Kachman, S. D., Spangler, M. L., Fernandes, S. B., et al. (2018). Efficiency of multi-trait, indirect, and trait-assisted genomic selection for improvement of biomass sorghum. genetica TAG. theoretical and applied genetics. *Theoretische und angewandte Genetik* 131, 747–755. doi: 10.1007/s00122-017-3033-y

Belamkar, V., Guttieri, M. J., Hussain, W., Jarquin, D., El-Basyoni, I., Poland, J., et al. (2018). Genomic selection in preliminary yield trials in a winter wheat breeding program. *G3: Genes Genomes Genet.* 8, 2735–2747. doi: 10.1534/g3.118.200415

Blondel, M., Onogi, A., Iwata, H., and Ueda, N. (2015). A ranking approach to genomic selection. *PLoS One* 10, e0128570. doi: 10.1371/journal.pone.0128570

Cerrudo, D., Cao, S., Yuan, Y., Martinez, C., Suarez, E. A., Babu, R., et al. (2018). Genomic selection outperforms marker assisted selection for grain yield and physiological traits in a maize doubled haploid population across water treatments. *Front. Plant Sci.* 9, 366. doi: 10.3389/fpls.2018.00366

Cheng, Q., Jiang, S., Xu, F., Wang, Q., Xiao, Y., Zhang, R., et al. (2021). Genome optimization *via* virtual simulation to accelerate maize hybrid breeding. *Briefings Bioinf.* 23, bbab447. doi: 10.1093/bib/bbab447

Collard, B. C., and Mackill, D. J. (2008). Marker-assisted selection: An approach for precision plant breeding in the twenty-first century. *Philos. Trans. R. Soc. B: Biol. Sci.* 363, 557–572. doi: 10.1098/rstb.2007.2170

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., et al. (2017). Genomic selection in plant breeding: Methods, models, and perspectives. *Trends Plant Sci.* 22, 961–975. doi: 10.1016/j.tplants.2017.08.011

de los Campos, G., and Pérez, P. (2015). "BGLR: Bayesian generalized linear regression," in *R package version 1.0*, 4.

Desta, Z. A., and Ortiz, R. (2014). Genomic selection: Genome-wide prediction in plant improvement. *Trends Plant Sci.* 19, 592–601. doi: 10.1016/j.tplants.2014.05.006

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with r package rrBLUP. *Plant Genome* 4, 250–255. doi: 10.3835/plantgenome2011.08.0024

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models *via* coordinate descent. *J. Stat. software* 33, 1. doi: 10.18637/jss.v033.i01

González-Recio, O., and Forni, S. (2011). Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. *Genet. Sel. Evol.* 43, 21329522. doi: 10.1186/1297-9686-43-7

Guo, Z., Wang, H., Tao, J., Ren, Y., Xu, C., Wu, K., et al. (2019). Development of multiple SNP marker panels affordable to breeders through genotyping by target sequencing (GBTS) in maize. *Mol. Breed.* 39, 1–12. doi: 10.1007/s11032-019-0940-4

Heffner, E. L., Lorenz, A. J., Jannink, J., and Sorrells, M. E. (2010). Plant breeding with genomic selection: gain per unit time and cost. *Crop Sci.* 50, 1681–1690. doi: 10.2135/cropsci2009.11.0662

Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. *R News* 2, 18–22.

Liu, H., Wang, X., Xiao, Y., Luo, J., Qiao, F., Yang, W., et al. (2020). CUBIC: An atlas of genetic architecture promises directed maize improvement. *Genome Biol.* 21, 1–17. doi: 10.1186/s13059-020-1930-x

Ma, W., Qiu, Z., Song, J., Li, J., Cheng, Q., Zhai, J., et al. (2018). A deep convolutional neural network approach for predicting phenotypes from genotypes. *Planta* 248, 1307–1318. doi: 10.1007/s00425-018-2976-9

McMurdie, P. J., and Holmes, S. (2015). Shiny-phyloseq: Web application for interactive microbiome analysis with provenance tracking. *Bioinformatics* 31, 282–283. doi: 10.1093/bioinformatics/btu616

Meuwissen, T. H., Hayes, B. J., and Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829. doi: 10.1093/genetics/157.4.1819

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2014). "Misc functions of the department of statistics (e1071), TU wien," in *R package version 1.6-3*.

Ornella, L., Pérez, P., Tapia, E., González-Camacho, J., Burgueño, J., Zhang, X., et al. (2014). Genomic-enabled prediction with classification algorithms. *Heredity* 112, 616–626. doi: 10.1038/hdy.2013.144

Qiu, Z., Cheng, Q., Song, J., Tang, Y., and Ma, C. (2016). *Application of machine learning-based classification to genomic selection and performance improvement* (Cham, Switzerland: Springer), 412–421.

Robert, P., Auzanneau, J., Goudemand, E., Oury, F.-X., Rolland, B., Heumez, E., et al. (2022). Phenomic selection in wheat breeding: Identification and optimisation of factors influencing prediction accuracy and comparison to genomic selection. *Theor. Appl. Genet.* 135, 895–914. doi: 10.1007/s00122-021-04005-8

Sievert, C. (2020). *Interactive web-based data visualization with r, plotly, and shiny* (Boca Raton, Florida: CRC Press).

Tanaka, R., and Iwata, H. (2018). Bayesian Optimization for genomic selection: A method for discovering the best genotype among a large number of candidates. *Theor. Appl. Genet.* 131, 93–105. doi: 10.1007/s00122-017-2988-z

Wang, S., Xu, Y., Qu, H., Cui, Y., Li, R., Chater, J. M., et al. (2021). Boosting predictabilities of agronomic traits in rice using bivariate genomic selection. *Briefings Bioinf.* 22, bbaa103. doi: 10.1093/bib/bbaa103

Xiao, Y., Jiang, S., Cheng, Q., Wang, X., Yan, J., Zhang, R., et al. (2021). The genetic mechanism of heterosis utilization in maize improvement. *Genome Biol.* 22, 1–29. doi: 10.1186/s13059-021-02370-7

Xu, Y., and Crouch, J. H. (2008). Marker-assisted selection in plant breeding: From publications to practice. *Crop Sci.* 48, 391–407. doi: 10.2135/cropsci2007.04.0191

Yan, J., and Wang, X. (2022). Unsupervised and semi-supervised learning: The next frontier in machine learning for plant systems biology. *Plant J.* 111, 1527–1538. doi: 10.1111/tpj.15905

Yan, J., Xu, Y., Cheng, Q., Jiang, S., Wang, Q., Xiao, Y., et al. (2021). LightGBM: Accelerated genomically designed crop breeding through ensemble learning. *Genome Biol.* 22, 1–24. doi: 10.1186/s13059-021-02492-y

Yu, Y., Ouyang, Y., and Yao, W. (2018). shinyCircos: An R/Shiny application for interactive creation of circos plot. *Bioinformatics* 34, 1229–1231. doi: 10.1093/bioinformatics/btx763