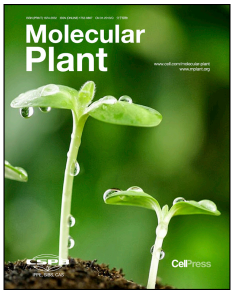


Smart Breeding Platform: a web-based tool for high-throughput population genetics, phenomics, and genomic selection

Huihui Li, Xin Li, Peng Zhang, Yingwei Feng, Junri Mi, Shang Gao, Lele Sheng, Mohsin Ali, Zikun Yang, Liang Li, Wei Fang, Wensheng Wang, Qian Qian, Fei Gu, Wenbin Zhou

PII: S1674-2052(24)00075-3

DOI: https://doi.org/10.1016/j.molp.2024.03.002

Reference: MOLP 1700

To appear in: *MOLECULAR PLANT*

Received Date: 25 September 2023

Revised Date: 4 February 2024

Accepted Date: 4 March 2024

Smart Breeding Platform: a web-based tool for high-throughput population genetics, phenomics, and genomic selection

Huihui Li[1,2,#,*], Xin Li[3,4,#], Peng Zhang[3,4], Yingwei Feng[1,2], Junri Mi[3,4], Shang Gao[1,2], Lele Sheng[3,4], Mohsin Ali[1,2], Zikun Yang[3,4], Liang Li[1], Wei Fang[1], Wensheng Wang[1,2], Qian Qian[1,2,5,6], Fei Gu[3,4,*], Wenbin Zhou[1]

[1] State Key Laboratory of Crop Gene Resources and Breeding, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences (CAAS), Beijing, China

[2] Nanfan Research Institute, CAAS, Sanya, Hainan, China

[3] DAMO Academy, Alibaba Group, Hangzhou, China

[4] Hupan Lab, Hangzhou, China

[5] Yazhouwan National Laboratory, No. 8 Huanjin Road, Yazhou District, Sanya City, Hainan Province 572024, China

[6] State Key Laboratory of Rice Biology, China National Rice Research Institute, Hangzhou, Zhejiang, China

* Corresponding author: Huihui Li, email: lihuihui@caas.cn

Fei Gu, email: gufei.gf@alibaba-inc.com

# These authors contributed equally

Runing title: Smart Breeding Platform

In the era of big data and artificial intelligence, "smart breeding" has become a broad conceptual framework encompassing the paradigm shift of crop breeding to relying on analysis of high-throughput population genetics and phenomics data to conduct genomic selection, allowing identification and optimal use of the genetic potential in crop species (Sharma et al., 2022; Xiao et al., 2022; Xu et al., 2022). Most existing tools for analyzing high-throughput breeding data require extensive computational power, complex installation processes, and command-line expertise, and are therefore challenging and inconvenient for the majority of researchers and breeders (Brandies and Hogg, 2021). To overcome these limitations, we developed Smart Breeding Platform (https://sbp.ibreed.cn), a user-friendly, web-based tool for management and analysis of large-scale genetic, genomic, and phenomic data. This platform is freely accessible through the internet and allows users to import data, perform various statistical analyses, and conduct genome-wide association studies and genomic selection using both classical machine learning and deep-learning models. It will enable plant breeders to easily conduct the following steps: (1) efficiently record, manage, and process raw phenotypic and genotypic data; (2) perform phenotypic and population genetic analyses in highly customizable ways; and (3) easily conduct GWAS and genomic selection using classical machine learning and deep-learning models. Smart Breeding Platform contains four main sections (Figure 1): (1) Germplasm Data Management, (2) Test Management, (3) Genomic Data Management, and (4) Data Analysis. Each section is described in detail below.

Germplasm Data Management

This section contains tables for storing germplasm and intermediate material data and pedigree data. The tables are directly editable and function similarly to a standard spreadsheet.

Germplasm and intermediate material table

In the germplasm and intermediate material table, germplasm metadata can be entered in a single row each, with the following parameters (columns) provided by default: germplasm name, year of seed production, breeding station, storage location, quantity harvested, quantity currently available, and serial number. Additional parameters, including images, can be added and customized by the user. The data in each row are directly exportable. Advanced lines can be promoted, allowing users to track advancement choices over time. A summary bar graph shows the number of plant lines across years, generations, and storage locations. All tables support efficient row-level filtering and sorting for quick data retrieval.

Pedigree module

The pedigree module contains pedigree records for existing germplasm. Information about parents and offspring can be viewed in either table or graph form. Graphs can be used to visualize the lineage of one or more lines, including inbred lines, hybrids, or both.

Graphs can also be customized to show information for only parental lines or progeny. In instances involving more than one generation, the depth of the visualized pedigree can also be adjusted. The entire pedigree for a breeding program can be displayed as a network graph that highlights the most popular lines.

Location module

By default, this table includes the following fields: year, season, breeding station, location name, longitude, latitude, size (e.g., the number of rows in a field trial), and environmental factors such as maturity zone, soil characteristics, and agronomic practices. These data can be used to compare variables between sites and track relevant factors that may contribute to phenotypic outcomes.

Warehouse in–out module

This module is used to track all seeds that enter and exit a specific research station. Users can log events such as seed allocation for yield trials, seed transfers to other breeders, or receipt of new germplasm. Relevant data including seed quantity and the time and date of transfers can be included. A bar chart allows users to examine changes in seed stocks over time.

In summary, the Germplasm Data Management section allows for efficient and intuitive storage and analysis of metadata for all germplasm used in a breeding program.

Test Management

Broadly, this section stores data about all field testing and nursery locations. Details of the two specific modules, the field testing module and the crossing nursery module, are included below.

Field testing module

This section includes tools for management of field testing. It contains functionality for specifying relevant lines, experimental designs, replicate numbers, and field layouts. Seven commonly-used experimental designs are included: completely randomized design, randomized complete block design (RCBD), augmented design, spatial design, sparse design, alpha-lattice design, and row-column design. After a design is selected by the user, the module can be used to automatically generate a suitable field layout based on the number of entries in the experiment and the number of available field plots. By including information about the physical location of a trial, multiple experiments from different breeders can be automatically placed within a single field. The field layout is output as a table containing the coordinates of each plot within the field. This module also includes heatmaps, which show the distribution of values for each trait across the field; a stability analysis, which shows the performance of specific lines across locations; and a testing history, demonstrating the trials and locations in which a specific line has been tested.

Crossing nursery

The crossing nursery module can be used to plan new lines and pedigrees. The user selects sets of female and male parental lines. The module then generates a crossing matrix, with options for user input regarding specific cross combinations (e.g., crossing patterns) and harvest instructions (for each row or plant). The module auto-generates inventory entries to be added to the germplasm table and adds the pedigree of each cross combination to the pedigree record table.

The Test Management section has tools to track experimental locations and plant research materials (i.e., seeds) with ease. It allows researchers to easily visualize available stock and to plan field experiments and crosses. Intuitive organization of these resources in a single location enables researchers to focus on planning and conducting high-level experiments.

Genomic Data Management

In this section, users can easily upload and manage all genomic sequencing data, reference genome files, and genomic variant files. The data stored in this section can then be used in the Data Analysis section.

Data Analysis

Phenotypic statistical analysis module

This module is used for analyzing high-throughput phenotypic data collected in the field. Multi-year, multi-location data can be extracted directly from the field testing module or can be uploaded separately. Based on the experimental design, the module can be used to fit a mixed linear model (MLM) to calculate best linear unbiased estimation (BLUE) or best linear unbiased prediction (BLUP) (Bates et al., 2015). The model fits two-dimensional spatial patterns for spatial designs to account for soil heterogeneity (Covarrubias-Pazaran, 2016; Rodriguez-Alvarez et al., 2018). Data can be analyzed separately for each location or as an integrated dataset including points from all locations. For each genotype, the module outputs the BLUP and BLUE values of the included traits. Variance components, heritability, and trait correlations can also be calculated. Entry-mean heritability and plot-mean heritability of each trait are derived from the variance components of random models. These two metrics can help breeders to assess the precision of trait values both at single-plot level and across locations.

This module also automatically calculates correlations for all pairs of phenotypic traits. The phenotypic correlation between each pair of traits is calculated as the Pearson correlation coefficient of the raw phenotypic data, whereas the genetic correlation between each pair of traits is calculated as the correlation of genetic effects in a model fitting both traits and residual correlation effects (Muñoz and Sanchez, 2020). Examination of trait correlations enables breeders to identify traits that can be bred independently (i.e., traits that have low correlations with other traits) and traits that must be separated or bred jointly (i.e., traits that have strong positive or negative correlations with other traits). For each linear model, the goodness of fit and the validity of the residual normality assumption can be assessed using diagnostic plots, including raw data distribution histograms, residual histograms, plots showing residual compared to fitted values, and residual Q–Q plots. Breeders can then select the best lines (those with ideal values across traits) using scatterplots that display the distribution and correlation of BLUP or BLUE values for pairs of traits. Overall, this module includes advanced single-trait and multi-trait analyses that can be conducted in an automated, user-friendly manner.

Genetic variation analysis module

This module can be used to efficiently identify genetic variants based on high-throughput genome sequencing data. In comparison to the standard pipeline for sequence alignment and germline variant-calling analysis (BWA+GATK) (Yin et al., 2021), the analysis method used here is significantly faster on our platform, due to the boosted tools with novel acceleration algorithm on the NVIDIA CUDA platform. Results of the new method are highly consistent with the standard BWA+GATK pipeline (99.9% accuracy) and are completed ~100× faster when two NVIDIA Turing T4 graphics cards are used. Inclusion of additional graphics cards would further improve the processing speed. The sequence alignment and sequencing depth can be visualized with Integrative Genomics Viewer (IGV) (Robinson et al., 2011), which has been optimized to load large genome dataset.

Genomic statistical analysis module

This module facilitates analyses of genetic diversity for a specific population. It takes SNP data as input, either as VCF files produced by the variant-calling module or as user-uploaded HapMap or VCF files. The module outputs some or all of the following 10 population genetics measures as specified by the user: allele frequency values, genotype frequency values, population divergence (*F*st) values, nucleotide diversity values, population structure results, a kinship matrix, a neighbor-joining tree, unweighted pair group method with arithmetic mean (UPGMA) clustering results, linkage disequilibrium (LD) values (r^2, D, and D'), and an LD graph. These analyses enable breeders to evaluate germplasm diversity and select the best lines for future crosses to maintain long term genetic gain. For example, the neighbor-joining tree (Paradis and Schliep, 2019) and UPGMA clustering show the genetic similarities among individuals in the population and enable breeders to assess the genetic diversity in the population.

GWAS analysis module

The GWAS module implements the 'GAPIT' R package (Wang and Zhang, 2021) to identify SNPs underlying phenotypic variations. Phenotype and marker data can be transferred directly from other modules in the platform or can be uploaded individually by the user. Seven models from the 'GAPIT' package are included: Generalized Linear Model (GLM), MLM, Compressed Mixed Linear Model (CMLM), Multi-Locus Mixed Model (MLMM), Settlement of MLM Under Progressively Exclusive Relationship (SUPER), Fixed and Random Model Circulating Probability Unification (FarmCPU), and Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway (BLINK). The user can define the minor allele frequency (MAF) for filtering SNPs and the number of principal components (PCs) to include in the model as fixed effects. The output includes genome-wide and single-chromosome Manhattan plots. Manhattan plots can also be generated for multiple traits to determine whether nearby SNPs control differing traits. Q–Q plots show the observed compared to the expected (i.e., uniformly-distributed) *p*-values. Additional summary graphs show the distributions of traits and markers in the genome and the LD values between nearby markers. Statistically significant markers are considered strong candidates for marker-assisted selection (MAS) for desired traits or for fine-mapping to identify causal genes for a specific phenotype.

Genomic selection analysis module

The genomic selection module is used to predict trait values among inbred lines, hybrids, or progenies based on molecular markers. It consists of three steps: dataset formation, training, and prediction. To generate a training dataset, the phenotype and marker data are uploaded or retrieved from the phenotype analysis module and the variant-calling module. In the dataset formation step, the program calculates the number of samples with both phenotype and marker data. In the model training step, the user selects a dataset, trait(s) of interest, and a model. The latter is either a statistical model such as genomic BLUP (GBLUP) (VanRaden, 2008; Endelman, 2011), a classical machine learning model, or a deep-learning model. After the model is trained, cross-validation is performed and the predictive accuracy is displayed. The user can then choose a trained model (e.g., the model with the highest predictive accuracy) to predict the performance of offspring from a cross or of the corresponding parental lines. This module yields predicted trait values for each specified line, allowing a breeder to select lines that are predicted to have optimal performance and to discard lines with undesirable traits. The breeder can thus select the most promising potential crosses from many possible combinations, saving time and resources. The predicted high performance inbreds or hybrids can be directly exported to a germplasm table for crossing or field evaluation.

Case study

To showcase the Smart Breeding Platform's capabilities, we utilized a rice dataset (Wang et al., 2018) with 100 varieties, each featuring multi-year, multi-location phenotypic data (Supplemental Table S1). Germplasm data were uploaded to the "rice100" table in the Germplasm Management module. The first five varieties advanced to "Advancement2023." A pedigree table simulated 35 records from ERS470485 and ERS470543. Three testing locations were added to the Location Management table, each with 20 rows and 20 ranges (totaling 400 plots). Field testing experiments, rice_2022 and rice_2023, followed RCBD designs across different locations. Phenotypic data underwent analysis in the Phenotypic Statistical Analysis module, generating BLUP and BLUE values, assessing genetic and phenotypic variance, heritability, and trait correlations.

In the Crossing Nursery module, 10 female and 10 male lines produced 28 two-way crosses. Germplasm data populated the 'Nursery23' table, and pedigree records filled the respective table. The Genomic Data Management module received the Os-Nipponbare-Reference-IRGSP-1.0 file and paired-end sequencing data. The Genetic Variation Analysis module conducted variant calling; 100 VCF files merged with criteria (depth > 50, quality value > 500), resulting in 55,589 SNPs. The Genomic Statistical Analysis module used SNPs for diverse population results. BLUP and BLUE values, along with merged marker data, identified significant SNPs in the GWAS module for plant height and heading date. The same datasets predicted values for new inbred and hybrid lines in the Genomic Selection module. Details of the analyses and results can be found in the platform manual. All data and results can be viewed by clicking “Enter as Guest” button on the login page.

In conclusion， this novel intelligent breeding platform integrates numerous data types (seed inventory, field testing, phenotypes, SNP markers, and plant crosses) with key analyses (GWAS, population genetic parameters, and genomic selection) in a single seamless system. All analytical tools have user-friendly interfaces and are simple to configure and run. The computing speeds for the genomic data analyses are substantially faster in this platform than in conventional tools. Smart Breeding Platform provides a comprehensive tool for the storage and management of germplasm data, experiments, and statistical analyses, allowing breeders to more easily identify and generate optimal germplasm, ultimately increasing the speed of genetic gain.

Figure

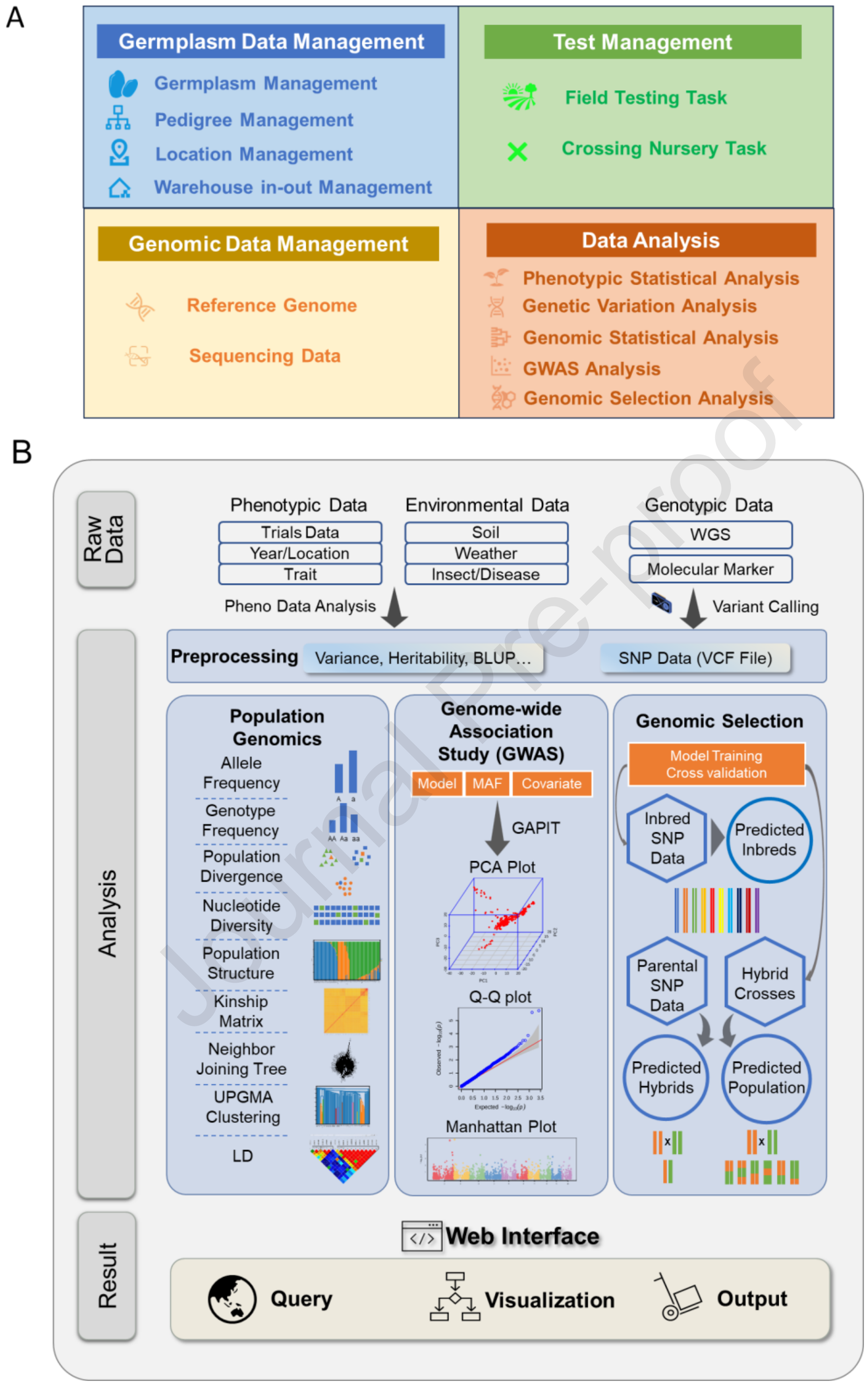


Figure 1. Main sections (A) and analysis workflow (B) of the Smart Breeding Platform. WGS, whole genome sequencing; BLUP, best linear unbiased prediction; SNP, single nucleotide polymorphism; UPGMA, unweighted pair group method with arithmetic mean; LD, linkage disequilibrium; and PCA, principal component analysis.

Funding

This work was supported by the Alibaba Foundation, the National Natural Science Foundation of China (32188102, 32361143514), Innovation Program of Chinese Academy of Agricultural Sciences, and the Project of Hainan Yazhou Bay Seed Lab (B21HJ0223).

Author contributions

W.Z., Q.Q., H.L., and F.G. conceived the project. X.L., P.Z., Y. F., J.M., S.G., L.S., M.A., H.L., and F.G. conducted data analyses and platform development. Z.Y. organized data used in case study. W.Z., Q.Q., L.L., W.W., and W.F. provided data and technical guidance. H.L., X.L., P.Z., M.A., and F.G. wrote the manuscript. All authors read and approved the manuscript.

Acknowledgements

The authors declare that they have no conflict of interest.

References

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. J. Stat. Softw. **67**:1-48. https://doi.org/10.18637/jss.v067.i01.

Brandies, P.A., and Hogg, C.J. (2021). Ten simple rules for getting started with command-line bioinformatics. Public Library of Science San Francisco, CA USA. https://doi.org/10.1371/journal.pcbi.1008645

Covarrubias-Pazaran, G. (2016). Genome-Assisted Prediction of Quantitative Traits Using the R Package sommer. PLoS One **11**:e0156744. 10.1371/journal.pone.0156744.

Endelman, J.B. (2011). Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. The Plant Genome **4**:250-255 https://doi.org/10.3835/plantgenome2011.08.0024.

Muñoz, F., and Sanchez, L. (2020). breedR: Statistical Methods for Forest Genetic Resources Analysts. R package version 0.12-5.

Paradis, E., and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics **35**:526-528. https://doi.org/10.1093/bioinformatics/bty633

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. **29**:24-26. https://doi.org/10.1038/nbt.1754

Rodriguez-Alvarez, M.X., Boer, M.P., van Eeuwijk, F.A., and Eilers, P.H. (2018). Correcting for spatial heterogeneity in plant breeding experiments with P-splines. Spatial. Stat. **23**:52-71. https://doi.org/10.1016/j.spasta.2017.10.003

Sharma, S., Kumar, K., Tribhuvan, K., Rita, Kumar, S., Jain, P., Saxena, S., Vijayan, J., Srivastava, H., and Gaikwad, K. (2022). High-throughput Genotyping Platforms. Genotyping by Sequencing for Crop Improvement. New York: John Wiley & Sons 22-37. https://doi.org/10.1002/9781119745686.ch2

VanRaden, P.M. (2008). Efficient methods to compute genomic predictions. J. Dairy Sci. **91**:4414-4423. https://doi.org/10.3168/jds.2007-0980

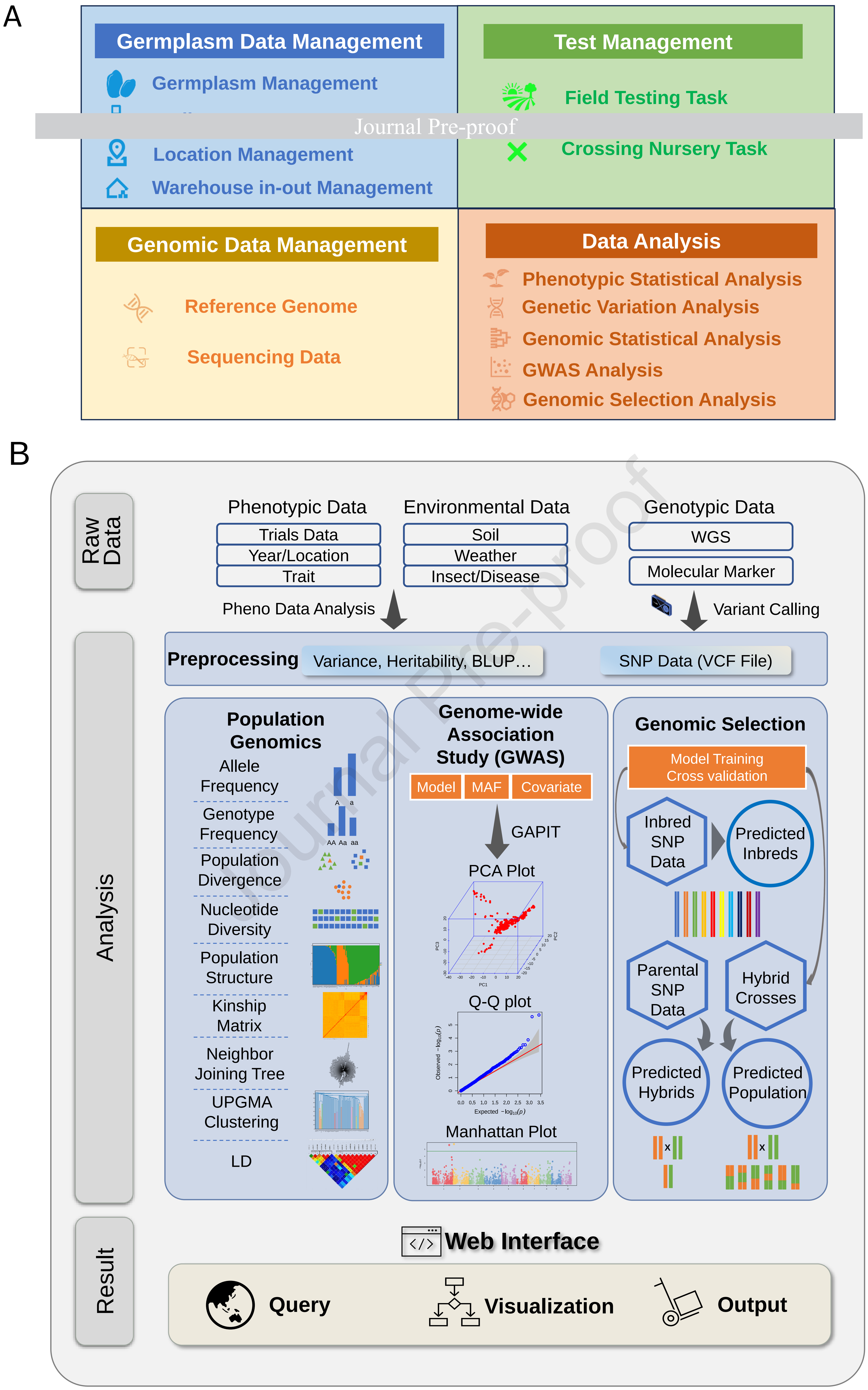
Wang, J., and Zhang, Z. (2021). GAPIT version 3: boosting power and accuracy for genomic association and prediction. Genomics Proteomics Bioinformatics **19**:629-640. https://doi.org/10.1016/j.gpb.2021.08.005

Wang, W., Mauleon, R. (2018). Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature **557**:43-49. https://doi.org/10.1038/s41586-018-0063-9

Xiao, Q., Bai, X., Zhang, C., and He, Y. (2022). Advanced high-throughput plant phenotyping techniques for genome-wide association studies: A review. J. Adv. Res. **35**:215-230. https://doi.org/10.1016/j.jare.2021.05.002

Xu, Y., Zhang, X., Li, H., Zheng, H., Zhang, J., Olsen, M.S., Varshney, R.K., Prasanna, B.M., and Qian, Q. (2022). Smart breeding driven by big data, artificial intelligence and integrated genomic-enviromic prediction. Mol. Plant **15**:1664-1695. https://doi.org/10.1016/j.molp.2022.09.001

Yin, L., Zhang, H., Tang, Z., Xu, J., Yin, D., Zhang, Z., Yuan, X., Zhu, M., Zhao, S., and Li, X. (2021). rMVP: a memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. Genomics Proteomics Bioinformatics **19**:619-628. https://doi.org/10.1016/j.gpb.2020.10.007


A
Germplasm Data Management
Germplasm Management
Location Management
Warehouse in-out Management
Test Management
Field Testing Task
Crossing Nursery Task
Genomic Data Management
Reference Genome
Sequencing Data
Data Analysis
Phenotypic Statistical Analysis
Genetic Variation Analysis
Genomic Statistical Analysis
GWAS Analysis
Genomic Selection Analysis
B
Raw Data
Phenotypic Data
Trials Data
Year/Location
Trait
Environmental Data
Soil
Weather
Insect/Disease
Genotypic Data
WGS
Molecular Marker
Pheno Data Analysis
Variant Calling
Analysis
Preprocessing
Variance, Heritability, BLUP…
SNP Data (VCF File)
Population Genomics
Allele Frequency
A a
Genotype Frequency
AA Aa aa
Population Divergence
Nucleotide Diversity
Population Structure
Kinship Matrix
Neighbor Joining Tree
UPGMA Clustering
LD
Genome-wide Association Study (GWAS)
Model
MAF
Covariate
GAPIT
PCA Plot
PC1
PC2
PC3
Q-Q plot
Expected −log10(p)
Observed −log10(p)
Manhattan Plot
Genomic Selection
Model Training
Cross validation
Inbred SNP Data
Predicted Inbreds
Parental SNP Data
Hybrid Crosses
Predicted Hybrids
Predicted Population
x
x
Result
Web Interface
Query
Visualization
Output