Che *et al. Plant Methods* (2025) 21:62
https://doi.org/10.1186/s13007-025-01383-8

Plant Methods

RESEARCH Open Access

Machine Learning-Based identification of resistance genes associated with sunflower broomrape

Yingxue Che[1], Congzi Zhang[1], Jixiang Xing[1], Qilemuge Xi[1], Ying Shao[2], Lingmin Zhao[2], Shuchun Guo[2*] and Yongchun Zuo[1*]

Abstract

Background Sunflowers (*Helianthus annuus* L.), a vital oil crop, are facing a severe challenge from broomrape (*Orobanche cumana*), a parasitic plant that seriously jeopardizes the growth and development of sunflowers, limits global production and leads to substantial economic losses, which urges the development of resistant sunflower varieties.

Results This study aims to identify resistance genes from a comprehensive transcriptomic profile of 103 sunflower varieties based on gene expression data and then constructs predictive models with the key resistant genes. The least absolute shrinkage and selection operator (LASSO) regression and random forest feature importance ranking method were used to identify resistance genes. These genes were considered as biomarkers in constructing machine learning models with Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Logistic Regression (LR), and Gaussian Naive Bayes (GaussianNB). The SVM model constructed with the 24 key genes selected by the LASSO method demonstrated high classification accuracy (0.9514) and a robust AUC value (0.9865), effectively distinguishing between resistant and susceptible varieties based on gene expression data. Furthermore, we discovered a correlation between key genes and differential metabolites, particularly jasmonic acid (JA).

Conclusion Our study highlights a novel perspective on screening sunflower varieties for broomrape resistance, which is anticipated to guide future biological research and breeding strategies.

Keywords Machine learning, Feature selection, Resistance genes, Sunflower broomrape

*Correspondence:
Shuchun Guo
200114050@163.com
Yongchun Zuo
yczuo@imu.edu.cn
[1]The State Key Laboratory of Reproductive Regulation and Breeding of Grassland Livestock, College of Life Sciences, Inner Mongolia University, Hohhot 010020, China
[2]Inner Mongolia Academy of Agricultural and Animal Husbandry Sciences, Hohhot 010000, China

Introduction

Sunflower (*Helianthus annuus* L.), one of the most extensively cultivated oil crops, faces severe threats from sunflower broomrape (*Orobanche cumana*). Sunflower broomrape is a holoparasitic angiosperm classified under the genus *Orobanche* in the *Orobanchaceae* family [1]. It parasitizes sunflower roots by connecting to the vascular tissues and absorbing water and nutrients from its host, severely impacting sunflower growth and even leading to plant death, which can result in serious yield losses [2, 3]. It is challenging to control sunflower broomrape in the field. However, the current preventive measures for sunflower broomrape remain quite restricted, focusing primarily on biological control [4], chemical control, physical control, and agricultural management strategies [5, 6]. Identifying resistance genes and developing resistant varieties are regarded as the most effective approaches to control broomrape [7, 8]. The genetic resistance of sunflower to broomrape is primarily mediated by major resistance genes [9, 10]. However, owing to the rapid evolution and genetic diversification within broomrape populations, novel virulent races continue to emerge. Consequently, this necessitates the ongoing discovery and characterization of new resistant gene candidates to maintain durable resistance in sunflowers [11, 12].

Machine learning is increasingly utilized in agriculture to assist with plant breeding and resistance screening [13–16]. The application of machine learning algorithms for Genomic selection (GS) breeding allows for the integration of genomic data and phenotypic information, enabling the prediction of phenotypes from genotypes [14, 17, 18]. It significantly improves the accuracy and efficiency of screening for resistant varieties [19]. The rapid progress of high-throughput omics technologies, such as transcriptomics and metabolomics, has led to the generation of larger, more complex, and noisier datasets [20, 21]. The heterogeneity and noise in these biological datasets compound the difficulty of analysis [22, 23]. Machine learning algorithms can process sophisticated datasets and identify key traits in plants, enabling more effective breeding programs and providing an efficient strategy for developing new varieties [24, 25].

Heike Sprenger et al. selected transcripts and metabolites from leaf samples of various potato varieties and successfully predicted drought tolerance in unstressed plants using a machine learning approach. This method simplifies the need for time-consuming and expensive drought stress experiments [26]. Xiaoxi Meng et al. trained a supervised classification model to identify genes that are transcriptionally responsive to cold stress and also performed cross-species prediction, which was accurate for cold-sensitive and cold-tolerant species. These results suggest that classifiers trained with stress data from well-studied species may be sufficient to predict gene expression patterns in related species with less sequenced genomes [27]. Wang et al. developed a machine learning model integrating multi-omics data to predict complex traits in Arabidopsis. The multi-omics approach significantly improved prediction accuracy and enabled systematic analysis of gene interactions [28]. Integration of multi-omics data with machine learning can help predict phenotypic traits, thereby accelerating the development of resistant cultivars [29–31].

In this study, we identified 24 potential resistance genes from gene expression data and utilized these genes to develop machine learning models to distinguish resistant and susceptible sunflower varieties based on gene expression data. Ultimately, the SVM model has excellent predictive performance and can accurately classify resistant and susceptible varieties. Furthermore, we found a correlation between some of the identified resistance genes and jasmonic acid (JA). Therefore, utilizing these methods for mining potential genes and predicting resistant variety holds significant reference value for agricultural breeding. These markers can facilitate trait prediction in unstressed plants, thereby eliminating the need for time-consuming and costly broomrape-stress experiments in the breeding process. The resistance genes identified in this study through machine learning methods are expected to guide the selection of resistant varieties.

Materials and methods

Plant materials

We selected 103 varieties of sunflower seeds (from the Inner Mongolia Academy of Agriculture and Animal Husbandry Sciences) for potted experiments. The mass of the growing medium and the mass of broomrape seeds were mixed according to the ratio of 1000 g:0.5 g. After adding nutrient soil, the growing medium-broomrape seed mixture, and nutrient soil to the pots in order, we sowed sunflower seeds to a depth of about 4 centimeters. Each variety comprised three replicates and was cultivated outdoors. When the sunflowers reached the early bud stage, their phenotypes were recorded, including plant height, the number of leaves, the number of broomrape emergences, fresh weight, and dry weight.

Sunflower varieties were classified according to the number of broomrape emergence. Sunflowers with a broomrape emergence count of 0 were classified as resistant varieties, and sunflowers with a broomrape emergence count of more than 0 were classified as susceptible varieties. One well-grown sunflower from each group was selected, and the sunflower roots were removed from the pots, washed with water, and then cut into centrifuge tubes using scissors and quickly placed in liquid nitrogen. The samples were processed for transcriptome sequencing. A subset of varieties (F059, F033, F087, F053, F096)

was selected for metabolism sequencing. Among them, F059 and F033 were resistant varieties and F087, F053, and F096 were susceptible varieties.

Transcriptome sequencing

According to the manufacturer's protocol, the purity, concentration, and integrity of RNA samples were examined by NanoDrop, Qubit 2.0, and Agilent 2100, respectively. High-quality RNA samples were selected for subsequent experiments. Subsequently, library construction was performed. Briefly, mRNA was isolated using Oligo(dT)-attached magnetic beads and then randomly fragmented in a fragmentation buffer. First-strand cDNA was synthesized using fragmented mRNA as a template. Second-strand synthesis was done with PCR buffer, dNTPs, RNase H, and DNA polymerase I. The cDNA was purified with AMPure XP, and double-strand cDNA was subjected to end repair. Adenosine was added to the end and ligated to adapters. The cDNA library was constructed through several rounds of PCR amplification.

Qubit 2.0 and Agilent 2100 were used to examine the concentration of cDNA and the insert size. Q-PCR was processed to obtain a more accurate and adequate library concentration.

Qualified libraries were uploaded to the Illumina NovaSeq 6000 platform for high-throughput sequencing with a read length of 150 bp paired.

Transcriptome analysis

The raw data quality was checked by FastQC (version 0.12.1). To ensure the quality and reliability of the data analysis, we filtered the raw data. First, the raw data were processed using Fastp (version 0.23.2) software [32], which included removing reads with N, removing low-quality reads, and removing the first 10 bp of the sequences. All subsequent analyses were performed based on high-quality, clean data. The sunflower reference genome (https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_002127325.2/) and gene annotation file were downloaded from NCBI (National Center for Biotechnology Information, https://www.ncbi.nlm.nih.gov/). The reference genome was indexed, and paired-end clean reads were aligned to the sunflower reference genome using HISAT2 (version 2.1.1) [33]. Finally, the. sam files were converted to.bam files by samtools (version 1.5) software [34]. The. bam files were aligned with the annotation files to quantify gene expression and obtain the gene expression matrix with FeaturesCounts (version 2.0.1) [35]. In the end, we obtained a gene expression matrix.

Differential expression analysis

The gene expression matrix was preprocessed, and genes with gene expression levels equal to 0 in more than 75% of the samples were filtered out. We performed the differential expression analysis on the gene expression matrix utilizing the edgeR package (version 4.0.12) [36]. Differentially expressed genes (DEGs) were considered significant at p. adjust < 0.05 and |log2Foldchange| > 1 [37]. Subsequently, to visually represent the differential expression results, the volcano plot was generated with the ggplot2 R package (version 3.4.4).

Weighted gene co-expression network analysis

Weighted gene co-expression network analysis (WGCNA) can potentially reveal gene networks and gene modules of biological significance related to broomrape resistance. Thus, we used the WGCNA package (version 1.72.5) [38] to perform WGCNA. Initially, the filtered gene expression matrix from the previous step was scaled through the varianceStabilizingTransformation function within the DESeq2 package (version 1.42.0) [39]. Then, the top 5,000 genes were selected for WGCNA based on the median absolute deviation (MAD) method. Subsequently, the goodSamplesGenes function of the WGCNA package [38] was employed to examine the unqualified samples.

After that, a scale-free co-expression gene network was constructed with the one-step network construction function in the WGCNA package. We selected an appropriate soft threshold power by the pickSoftThreshold function. Next, an adjacency matrix was constructed using the soft threshold power, and the adjacency matrix was transformed into a topological overlap matrix (TOM) [40]. Then, gene modules were divided using hierarchical clustering, with a module merging threshold set to 0.25, and the minimum module size was defined as 30. The modules associated with the resistant and susceptible traits were identified, and each module's eigengenes were obtained. Concurrently, the relationships between the modules and the traits were evaluated. Ultimately, the VennDiagram [41] tool was employed to ascertain the shared genes between the modules with the most significant correlations to traits and DEGs from the previous step. These shared genes were subjected to functional enrichment analysis.

The Gene Ontology (GO) [42] enrichment results were obtained from DAVID (https://david.ncifcrf.gov/) online tools [43, 44]. The Kyoto Encyclopedia of Genes and Genomes (KEGG) [45] pathway was enriched with the clusterprofiler (version 4.10.0) R package [46]. Then, the shared genes served as feature genes to construct machine learning models, including SVM, KNN, LR, and GaussianNB.

Model construction

The machine learning algorithms were implemented using the scikit-learn (version 1.3.2) library in Python

(version 3.11.5). The Python packages Numpy (version 1.26.0) and Pandas (version 2.1.3) were used to read and process the data.

Firstly, the filtered gene expression matrix was processed using the quartile method, and the processed gene expression data was standardized with the StandardScaler method in the scikit-learn library.

Then, leave-one-out cross-validation (LOOCV) partitions the dataset into training and test sets. LOOCV divided the datasets into n folds and then trained and evaluated the model n times, each time using one data point as the test set and the remaining n-1 data points as the training set [21, 47, 48]. To identify essential genes related to broomrape resistance, we utilized two feature selection methods: LASSO and random forest feature importance ranking. Based on LOOCV, the genes obtained from the feature selection method were used as input features to train machine learning models (KNN, LR, SVM, GaussianNB). Hyperparameter tuning was performed using the grid search. ROC curves were plotted using Matplotlib (version 3.8.1).

Model evaluation

The metrics for assessing the classification model are presented as follows [19]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$F1 - score = \frac{2 * (precision * recall)}{precision + recall} \quad (4)$$

TP, TN, FP, and FN stand for the numbers of true positives, true negatives, false positives, and false negatives, respectively. The area under the curve (AUC) was also used to evaluate the performance of the model [49].

Metabolites extraction

The LC/MS system for untargeted metabolomics analysis was conducted on the Waters Acquity I-Class PLUS ultra-high performance liquid chromatograph coupled with a Waters Xevo G2-XS QTof high-resolution mass spectrometer. The column employed was the Waters Acquity UPLC HSS T3 column (1.8 μm, 2.1 mm × 100 mm). Positive ion mode included mobile phase A (0.1% formic acid aqueous solution) and mobile phase B (0.1% formic acid acetonitrile). Negative ion mode included mobile phase A (0.1% formic acid aqueous solution) and mobile phase B (0.1% formic acid acetonitrile).

LC-MS/MS analysis

The Waters Xevo G2-XS QTOF high-resolution mass spectrometer enabled the acquisition of primary and secondary mass spectrometry data in MSe mode using MassLynx (version 4.2) software. Each acquisition cycle collected dual-channel data simultaneously at both low and high collision energies. The low collision energy was 2 V, while the high collision energy range was 10–40 V, and the scan frequency was 0.2 s for a mass spectrum. The parameters of the ESI ion source were as follows: capillary voltage: 2,000 V (positive ion mode) or -1,500 V (negative ion mode); cone voltage: 30 V; ion source temperature: 150 °C; desolvent gas temperature: 500 °C; backflush gas flow rate: 50 L/ h; desolvent gas flow rate: 800 L/h.

Data preprocessing and annotation

Progenesis QI software was used for peak extraction, peak alignment, and other data processing on the raw data collected using MassLynx. Identification was performed using the online METLIN database within Progenesis QI software.

Data analysis

The metabolomics data were aligned, normalized, log-transformed, and scaled using a mean-centering approach in MetaboAnalystR (version 4.0.0) [50]. Principal component analysis (PCA) assessed the variability between and within groups. Additionally, MetaboAnalystR was utilized to perform a t-test (student's t-test) and calculate fold change values. Differential metabolites were identified based on the criteria of |log2Foldchange| > 1 and p-value < 0.05.

Correlation of genes from the LASSO method and metabolism

We selected metabolomic data from six independent biological replicates and performed a correlation analysis with transcriptomic data. This analysis aimed to identify relationships between key genes identified through machine learning and differential metabolites. Using the cor function in the WGCNA package, we calculated Pearson correlation coefficients and p-values between key genes and differential metabolites. Finally, we constructed and visualized a correlation network in Cytoscape (version 3.10.2) [51] to refine and illustrate the interactions between genes and metabolites.

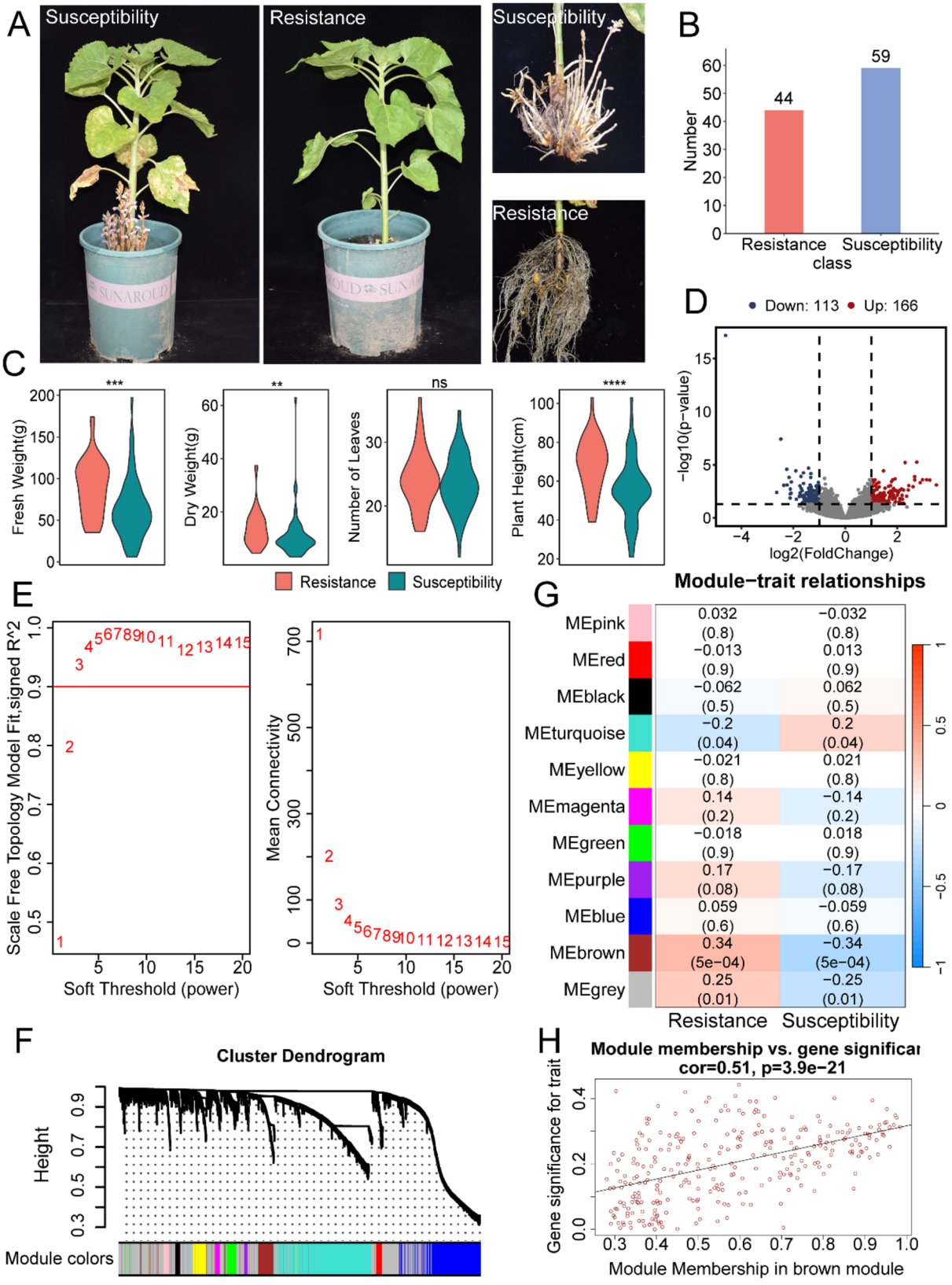

Fig. 1 (See legend on next page.)

(See figure on previous page.)

Fig. 1 Sunflower phenotypes and results of WGCNA analyses. **A** Schematic diagram of the phenotypes of resistant and susceptible varieties after infestation with broomrape and pictures of roots parasitized by broomrape. **B** Bar plot showing the number of susceptible and resistant varieties. **C** Plot of phenotypes of different classifications. **D** Volcano plot of differential genes of edgeR analysis. Blue represents the down-regulation of gene expression, and red represents the up-regulation of gene expression. **E** The scale-free fit index for various soft-thresholding powers (β) and the mean connectivity for various soft-thresholding powers. **F** Dendrogram of genes clustered via the dissimilarity measure. **G** Heatmap of the correlation between module eigengenes and traits. Grey module indicates genes that do not match any module. **H** Scatter plot between GS and MM in the brown module

Results

Phenotypes of sunflowers after being infested with Broomrape and gene co-expression network analysis

We selected 103 varieties of sunflower seeds for potted experiments. When the sunflowers reached the early bud stage, the number of broomrape emergences was recorded. The following criteria were used to classify resistant and susceptible varieties: sunflowers with a broomrape emergence number of 0 were classified as resistant varieties, and sunflowers with a broomrape emergence number of more than 0 were classified as susceptible (Fig. 1A). Then we got 44 resistant sunflower varieties and 59 susceptible sunflower varieties (Fig. 1B). To study the phenotypic changes in sunflower plants during the broomrape infection period, we recorded the phenotypic information of sunflower plants at the budding stage, including fresh weight, dry weight, number of leaves, and plant height. Compared to resistant plants, susceptible varieties exhibited a significant decrease in plant fresh weight, dry weight, and plant height after broomrape infestation of sunflower roots. However, there was no obvious change in the number of leaves (Fig. 1C).

To compare the differences between the resistant and susceptible varieties, we performed differential gene expression analysis. We eventually obtained 32,277 genes, and 279 differential expressed genes were screened, including 166 genes that were up-regulated and 113 genes that were down-regulated (Fig. 1D). WGCNA can potentially reveal gene networks and gene modules of biological significance related to broomrape resistance. Thus, WGCNA was performed. Using 0.9 as the threshold, we determined a soft threshold of 3 ($\beta = 3$) and constructed a scale-free network (Fig. 1E). Finally, 11 modules were identified based on average hierarchical clustering and dynamic tree clipping (Fig. 1F). The brown module was highly related to the resistance samples (298 genes, $r = 0.34$, $p = 5e\text{-}04$) (Fig. 1G). The scatter plot showed a correlation between gene significance (GS) and module membership (MM) in the brown module (cor = 0.51, $p = 3.9e\text{-}21$) (Fig. 1H).

Building machine learning models using genes screened by traditional biological methods

To better select for genes associated with biological traits, there were 32 shared genes obtained between the brown module genes and DEGs (Fig. 2A). The significant GO enrichment results of the shared genes were plasma membrane (GO:0005886) and transmembrane transport (GO:0055085) (Fig. 2B). It has been shown in the literature that invasive cells of broomrape penetrated the host's root cells, causing the host cell walls to disintegrate. Additionally, some cells experienced disruption to the plasma membrane, and degradation of the cytoplasm, until cell death [52]. This suggested that the plasma membrane played a crucial role in the infestation of sunflowers by broomrape. The enrichment analysis of the KEGG showed that the circadian rhythm pathway (han04712) was significantly enriched. Some literature shows that circadian rhythm may interfere with gene expression under stress [53].

To investigate whether the shared genes can accurately distinguish between resistant and susceptible sunflower varieties, we used the shared genes as the feature genes to construct the machine learning model, including SVM, KNN, LR, and GaussianNB to predict the resistance of broomrape for sunflowers. Then, we evaluated the performance of different models. The performance of the machine learning models was shown using the receiver operating characteristic curve (ROC) (Fig. 2C, D).

Construction of machine learning models and identification of resistance gene sets

Machine learning methods were used to select candidate genes for broomrape-resistant sunflower varieties. To identify the essential genes related to sunflower resistance, we used two feature selection methods for feature selection and evaluated the impact of different feature selection methods on model performance. First, we employed the LASSO regression method (Fig. 3A) to select resistance-related genes. When the lambda (λ) = 0.1, 24 resistance-related genes were selected (Fig. 3B, Table 1). Then, we constructed 4 machine learning models (SVM, LR, KNN, and GaussianNB) based on LOOCV with 24 identified genes. The result showed that the SVM model exhibited the greatest area under the curve (AUC) value in the test dataset (Fig. 3D) and also achieved a high AUC value in the training dataset (Fig. 3C), indicating strong classification performance.

Next, prediction models based on LOOCV were built with the random forest feature importance ranking (Fig. 4A). The importance scores for genes were calculated using the random forest feature importance method. We selected the top 10 genes in terms of feature importance scores. Subsequently, we constructed four

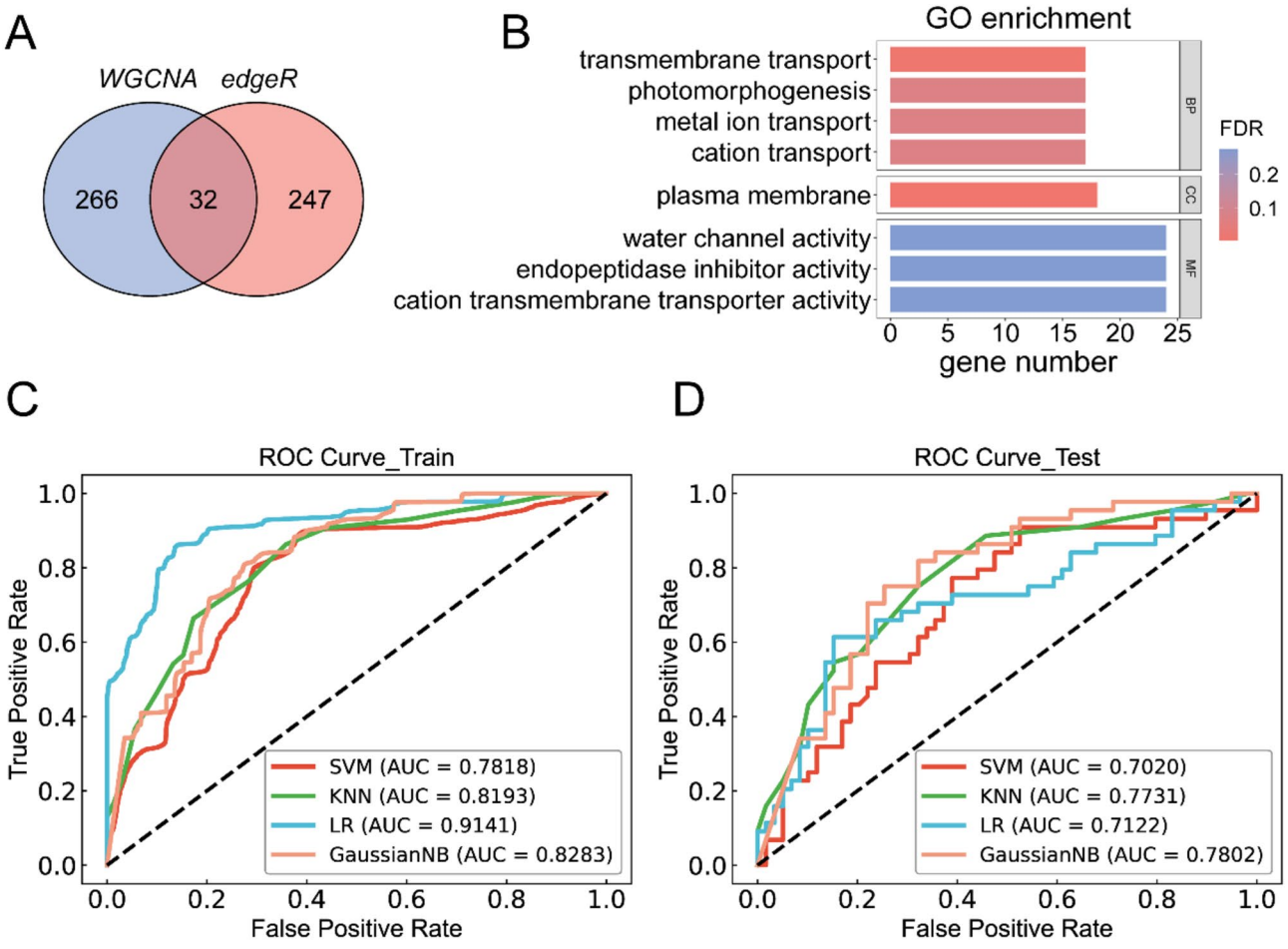

Fig. 2 Performance of models constructed using shared genes. **A** A total of 32 shared genes were identified between key module genes and DEGs. **B** The GO enrichment results from shared genes, GO terms with False Discovery Rate (FDR) < 0.05 were considered significant. **C** Model prediction performance of train dataset. **D** Model prediction performance of test dataset

machine learning models (SVM, LR, KNN, and GaussianNB) based on the top 10 genes. In the training and test datasets, the AUC values of most models exceeded 0.8. The LR model had the best performance (Fig. 4B, C). Next, we performed partial dependence plot (PDP) analysis on selected features using the LR model to examine how these features influence the model's predictions. The partial dependency plots were shown in Fig. 4D, E, and F. The results of the partial dependency plot analysis showed that the expression of the gene LOC110876215 was positively correlated with the probability of disease resistance, and the greater the expression of the gene, the higher the probability of disease resistance predicted by the model. LOC110927507 was negatively correlated with the expected probability of disease resistance. Figure 4F illustrated the bidirectional partial dependence of LOC110876215 and LOC110927507 in the LR model. The figure suggested that these two genes may exhibit a synergistic effect, collectively influencing the model's prediction outcomes.

Model performance evaluation

Comparing the AUC values of the models constructed from genes obtained by traditional biological methods and the AUC values of the models constructed from genes obtained by machine learning methods, the results showed that the accuracy of the models constructed from genes identified with traditional biological methods can be improved by machine learning methods to select resistance-related genes and construct models.

Afterward, we evaluated the performance of the models constructed from genes obtained by machine learning methods. The results showed that the 4 machine learning models constructed using the resistance-related genes screened by the LASSO method had the best performance. In the training dataset, the SVM model accuracy and AUC value reached 0.9523 (Fig. 5A) and 0.9917 (Fig. 3C), respectively. In the test dataset, the results showed that the SVM model achieved the most optimal prediction performance among the 4 machine learning models, achieving an accuracy, precision, recall, and

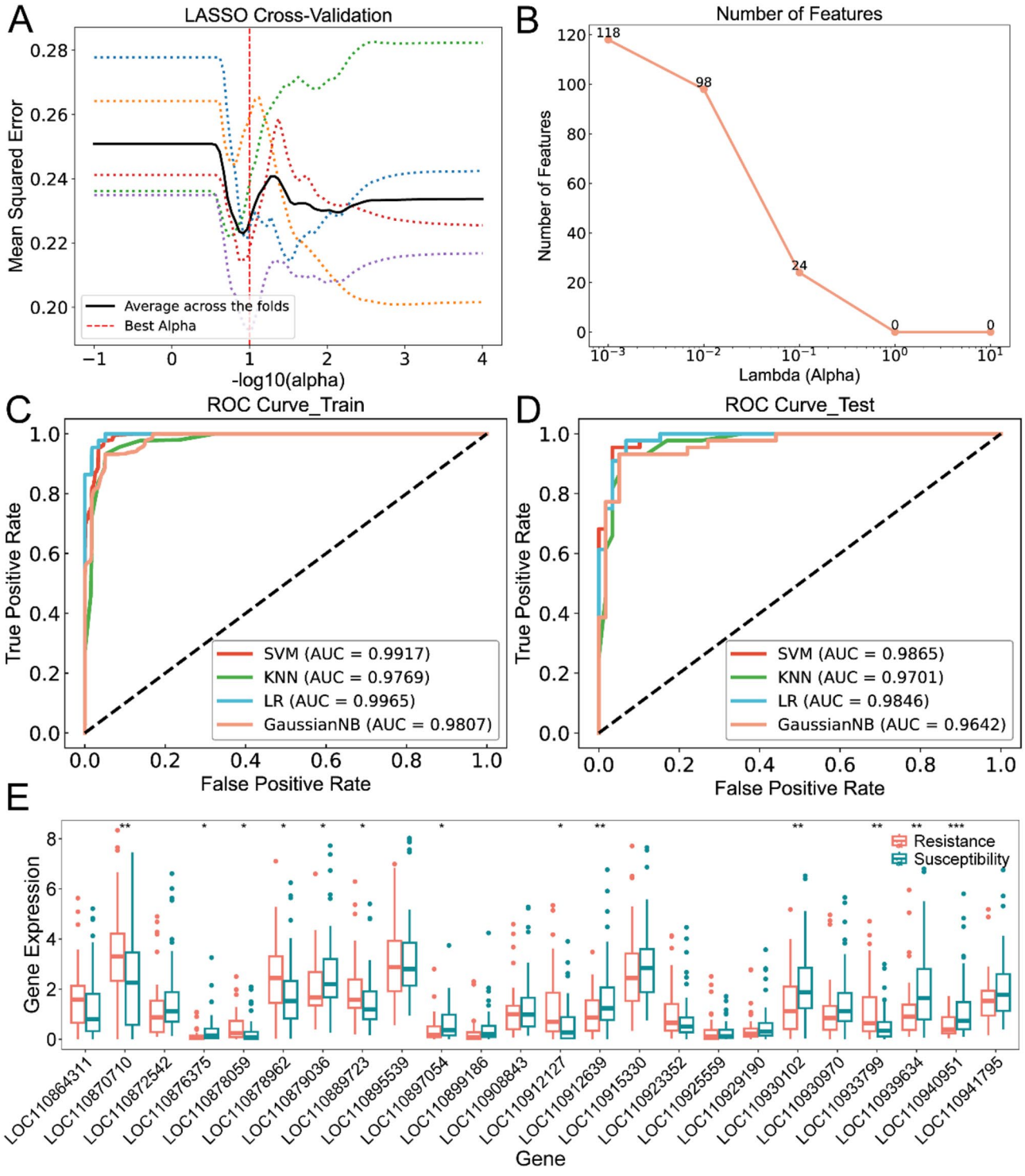

Fig. 3 Prediction model was constructed using key genes identified through the LASSO method. **A** LASSO regression algorithm. **B** Number of features for different lambda values. **C** Model prediction performance of training dataset. **D** Model prediction performance of test dataset. **E** Expression of 24 key genes in the two groups. *, **, and *** represent statistical significance at $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively

Table 1 The description of key genes from LASSO

Gene Symbol	Gene Description
LOC110872542	uncharacterized
LOC110895539	eukaryotic translation initiation factor 6-2
LOC110923352	uncharacterized
LOC110870710	putative UPF0481 protein At3g02645
LOC110912639	probable aquaporin TIP3-2
LOC110889723	protein transport protein SFT2
LOC110925559	uncharacterized
LOC110864311	protein NRT1/ PTR FAMILY 2.7
LOC110940951	glutathione S-transferase F11
LOC110908843	NADH kinase
LOC110899186	zinc finger CCCH domain-containing protein 20
LOC110941795	geranylgeranyl diphosphate reductase, chloroplastic
LOC110915330	putative glucose-6-phosphate 1-epimerase
LOC110912127	UDP-glycosyltransferase 75C1-like
LOC110897054	basic leucine zipper 43
LOC110878962	cytochrome P450 CYP82D47
LOC110878059	uncharacterized
LOC110876375	rhamnogalacturonate lyase B
LOC110930970	uncharacterized
LOC110939634	tRNA wybutosine-synthesizing protein 2/3/4
LOC110930102	beta-fructofuranosidase, insoluble isoenzyme CWINV3-like
LOC110929190	fatty alcohol: caffeoyl-CoA acyltransferase
LOC110879036	basic leucine zipper 43
LOC110933799	probable glycosyltransferase At3g07620

F1-score of 0.9514, 0.9535, 0.9318, and 0.9425 (Fig. 5B). The AUC of the SVM model was 0.9865 (Fig. 3D).

In conclusion, the results illustrated that the proposed SVM model exhibited excellent classification performance and accurately classified resistant and susceptible sunflower varieties.

Function of resistance genes and enrichment analysis

After the identification of 24 resistance-related genes through the LASSO approach, we examined a handful of genes that stand out for their potential impact, including LOC110899186 (zinc finger CCCH domain-containing protein 20), LOC110912127 (UDP-glycosyltransferase 75C1-like), and LOC110878962 (cytochrome P450 CYP82D47).

LOC110899186 encodes a protein characterized by a CCCH-type zinc finger domain. LOC110912127, a UDP-glycosyltransferase, exhibited higher expression levels in resistant varieties compared to susceptible ones. LOC110878962 showed increased expression in resistant sunflower varieties (Fig. 3E). It has been documented that the above genes are associated with stress tolerance [54–56]. Therefore, we speculated that LOC110899186, LOC110912127, and LOC110878962 were associated with sunflower resistance and may be involved in the resistance response to broomrape.

The 24 genes underwent enrichment analysis to determine the biological processes and pathways controlled by each gene. The GO enrichment result of the 24 genes was the carbohydrate metabolic process (GO:0005975), but it was not significant (p-value = 2.6e-2, p. adjust = 6.5e-1).

Correlation between resistance genes and differential metabolism

To systematically investigate the functions of genes identified through the LASSO method, correlations between these genes and differential metabolites were analyzed to explore their interactions. We first conducted a principal component analysis (PCA) on the metabolites. The first and second principal components (PC1 and PC2) explained 51.85% of the total variance, as demonstrated in the 2D score plots. PC1 was attributed a value of 32.98%, and PC2 was attributed a value of 18.87% (Fig. 6A). According to our criteria for differential metabolite expression analysis, we identified 692 differential metabolites, 435 were down-regulated and 257 were up-regulated (Fig. 6B). Differential metabolite heatmaps demonstrated that resistant and susceptible samples differed from each other (Fig. 6C).

Afterwards, we performed correlation analyses between differential metabolites and resistance genes. Ultimately, we obtained a network diagram of interactions between differential metabolites and resistance genes. As shown in Fig. 6D, genes identified using the LASSO method were associated with differential metabolites. Specifically, we found the genes LOC110915330 (putative glucose-6-phosphate 1-epimerase), LOC110941795 (geranylgeranyl diphosphate reductase, chloroplastic), and LOC110879036 (basic leucine zipper 43) were positively correlated with the metabolite JA (Fig. 6D).

Studies examining plant responses to parasitic plants have shown that both salicylates and jasmonates are crucial in mediating effective defenses [57]. We identified JA but did not detect salicylic acid (SA) in our differential metabolite analysis. This result suggested that jasmonic acid may regulate these genes, potentially contributing to sunflower resistance against broomrape invasion.

Discussion

This study established a comprehensive transcriptional profile of sunflowers, identified resistance genes based on gene expression data, and applied machine learning algorithms to classify the resistance and susceptibility of sunflowers to broomrape. However, gene expression datasets in plant research often exhibit high dimensionality and low sample sizes, which can lead to model overfitting and reduced generalization. Many features are redundant, affecting the performance of the model. Therefore, feature selection is an essential step before training ML

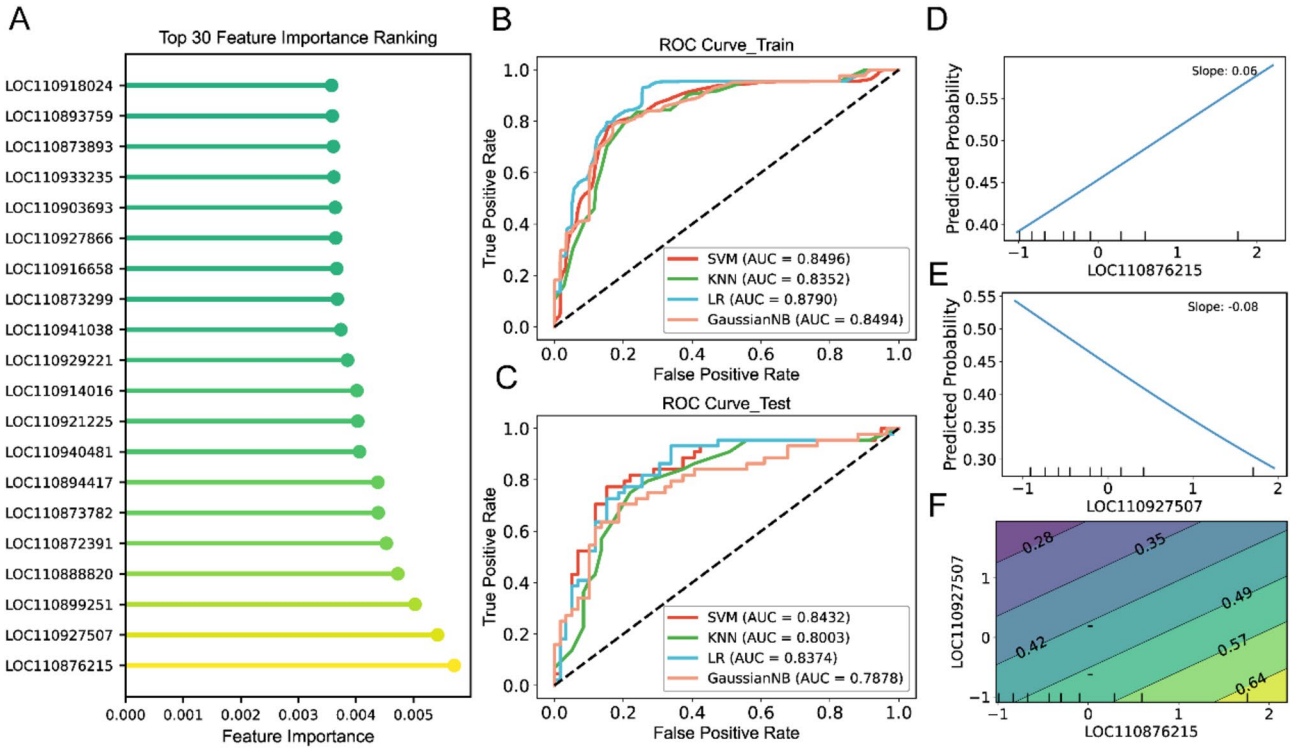

Fig. 4 Prediction model constructed by Random Forest importance rank. **A** The top 30 feature importance ranking. **B** Model prediction performance of training dataset. **C** Model prediction performance of test dataset. **D**, **E** Partial dependency graph of the top 2 features of the selected features. **F** The contour plot visually demonstrates the influence of the top 2 genes on the predicted values of LR

models. LASSO is commonly used for variable selection, especially in situations with a large number of predictors and few observations [58, 59]. It applies the L1-penalty (lambda) to shrink the coefficients of less important variables to zero, thereby filtering out non-essential features and constructing the optimal classification model.

In the end, 24 key genes were selected based on the LASSO regression and are listed in Table 1. Some of these genes are associated with biotic stress, while others are uncharacterized. LOC110899186 (zinc finger CCCH domain-containing protein 20) encodes a protein with a CCCH-type zinc finger domain. Zinc finger proteins represent a large family of transcriptional regulators in plants. Most zinc fingers are of the C2H2-type or CCCC-type and are involved in the regulation of plant growth, development, and stress adaptation [60]. Although CCCH-type zinc-finger proteins are relatively uncommon, they play a key role in regulating abiotic and biotic stress responses in plants [61, 62]. LOC110912127 (UDP-glycosyltransferase 75C1-like) is a UDP-glycosyltransferase. Previous studies have shown that the glycosyltransferase family plays a significant role in plant responses to abiotic stress, with glycosyltransferase family 1, also called UDP-glycosyltransferases (UGTs), being the most extensively studied group in plants [55]. Cytochrome P450s are a large family of protein-coding genes in plant genomes. LOC110878962 (cytochrome P450 CYP82D47) is one such P450 gene that plays a prominent role in responding to biotic stress [56].

The classification models discussed in this research included SVM, KNN, GaussianNB, and LR. KNN is a widely used supervised learning algorithm for classification tasks [63]. It predicts the category of a data point based on the categories of its k nearest neighbors by voting. LR regression is a supervised learning algorithm for binary classification problems that predicts the probability of an event occurring and is often used to determine the likelihood of an input belonging to a particular class. GaussianNB is a common machine learning classification algorithm for small samples with high dimensionality [64]. SVM is a regression and classification model. SVM aims to identify the best hyperplane to separate the samples into classes. However, SVM tends to overfit more easily than other algorithms [48].

We investigated the performance of the machine learning classifier. Ultimately, we constructed a prediction model (SVM) using 24 resistance genes to predict sunflower resistance to broomrape. SVM was the most accurate model for classifying sunflowers based on gene expression, with a high accuracy of 0.9514 and an AUC of 0.9865 in resistance and sensitivity for broomrape. Moreover, precision, recall, and F1 scores were implemented to evaluate the models. The model accurately categorized

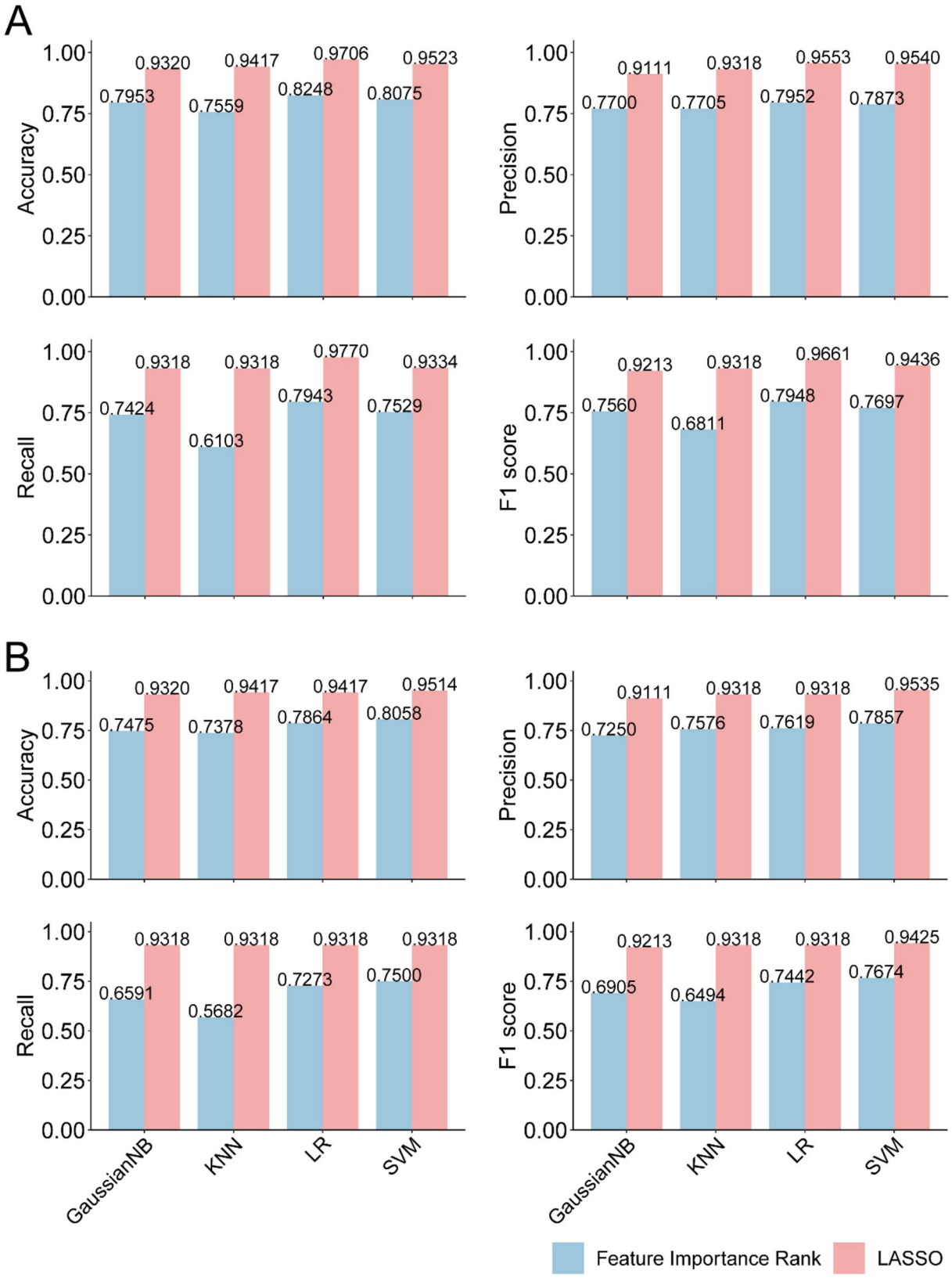

Fig. 5 The performance of the machine learning model. **A** The performance of the machine learning model on the training datasets. **B** The performance of the machine learning model on the test datasets

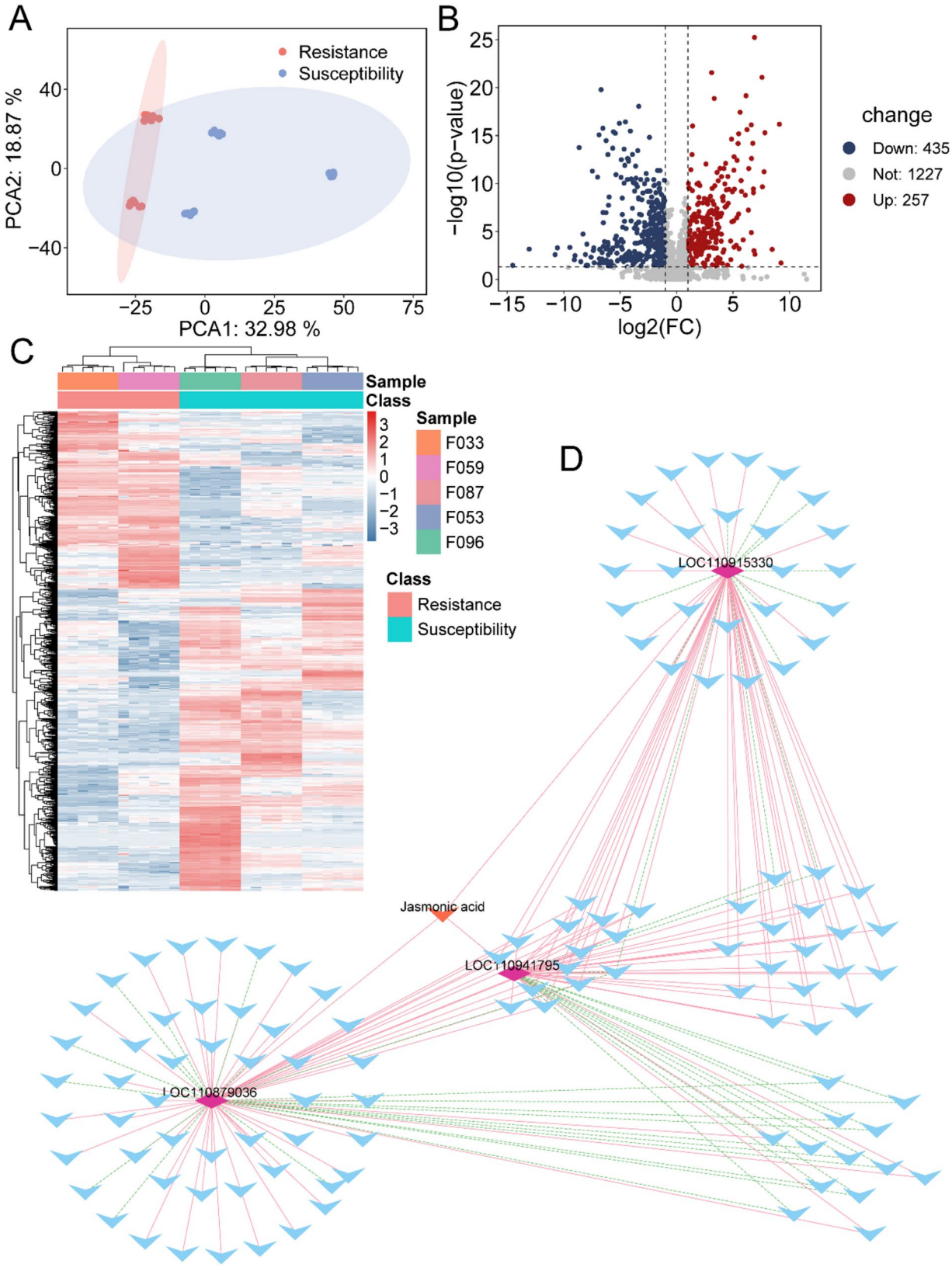

Fig. 6 The results of metabolic analysis. **A** PCA plot. **B** Volcano plot of differential metabolites. **C** Heatmap showed the results of the clustering analysis of differential metabolites. **D** Correlation between genes from LASSO and differential metabolites. Diamonds represented key genes from LASSO; red lines represented positive correlation; green lines represented negative correlation

the resistance and sensitivity varieties and produced excellent prediction results.

The results demonstrated that employing machine learning methods to select resistance-related genes and construct models could enhance the accuracy of models built using genes identified by traditional biological methods. It indicated that the machine learning methods could better capture the key genes in plants under stress than traditional biological methods. We hypothesize that these genes may be available to predict traits in unstressed experimental plants.

We found that the resistance genes selected by the LASSO method were associated with JA. JA is an important phytohormone, a signaling molecule that combats biotic and abiotic stresses [65]. When plants encounter adverse conditions, they activate stress responses that ultimately enhance their resistance and tolerance, improving their adaptability to environments. *Arabidopsis* responses to *Orobanche ramosa* involve changes in the expression of several JA-regulated genes [66]. Some studies suggest that plant parasitism is a biotic stress, and host responses to parasitic plants may involve mechanisms similar to those activated by other biotic and abiotic stresses [67]. For instance, Justin B. Runyon et al. discovered that plants, similar to their responses to herbivore and pathogen attacks, can detect parasitic plant invasion and activate induced defense mechanisms involving JA and SA signaling pathways [68].

In this study, we identified resistance genes of sunflower broomrape and developed a machine learning binary classification model to achieve accurate screening of antagonistic and perceptual varieties. Despite the strong classification performance, this research has some shortcomings, including the small sample size and the lack of an external dataset for model validation. This research demonstrated that machine learning algorithms could efficiently and accurately identify resistant key genes and build predictive disease resistance models to classify resistant varieties, eliminating the need for time-consuming and laborious traditional breeding experiments. The application of machine learning to biology to build predictive models for breeding is expected to guide future research efforts.

Acknowledgements
We thank the Inner Mongolia Academy of Agriculture and Animal Husbandry Sciences provided resources for performing these studies.

Author contributions
Y.Z. and S.G. conceptualized and supervised the project. Y.C. performed the data analysis and drew the figures. C.Z. recorded data and collected samples. J.X. reviewed the manuscript. Y.S. and L.Z. collected samples. Y.C. wrote the manuscript. Q.X. edited and reviewed the manuscript. All authors read and approved the manuscript.

Funding
This work was supported by the National Nature Scientific Foundation of China (No: 62171241, 62461046), the China Agriculture Research System of MOF and MARA (CARS-14-1-27), the Group Project of Developing Inner Mongolia through Talents (2025TEL25), and the Inner Mongolia Agricultural and Animal Husbandry Innovation Fund Project (2023CXJJN07).

Data availability
The data that supports the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 18 February 2025 / Accepted: 29 April 2025
Published online: 16 May 2025

References

1. Mutuku JM, Cui S, Yoshida S, Shirasu K. Orobanchaceae parasite-host interactions. New Phytol. 2021;230:46–59.
2. Louarn J, Boniface MC, Pouilly N, Velasco L, Pérez-Vich B, Vincourt P, Muños S. Sunflower resistance to Broomrape (Orobanche cumana) is controlled by specific QTLs for different parasitism stages. Front Plant Sci. 2016;7:590.
3. Xu Y, Zhang J, Ma C, Lei Y, Shen G, Jin J, Eaton DAR, Wu J. Comparative genomics of orobanchaceous species with different parasitic lifestyles reveals the origin and Stepwise evolution of plant parasitism. Mol Plant. 2022;15:1384–99.
4. Chabaud M, Auriac M-C, Boniface M-C, Delgrange S, Folletti T, Jardinaud M-F, Legendre A, Pérez-Vich B, Pouvreau J-B, Velasco L et al. Wild Helianthus species: A reservoir of resistance genes for sustainable pyramidal resistance to Broomrape in sunflower. Front Plant Sci. 2022;13:1038684.
5. Saabna N, Keasar T. Parasitoids for biological control in dryland agroecosystems. Curr Opin Insect Sci. 2024;64:101226.
6. Fernández-Aparicio M, Reboud X, Gibot-Leclerc S. Broomrape weeds. Underground mechanisms of parasitism and associated strategies for their control: A review. Front Plant Sci. 2016;7:135.
7. Fernández-Aparicio M, Del Moral L, Muños S, Velasco L, Pérez-Vich B. Genetic and physiological characterization of sunflower resistance provided by the wild-derived Or(Deb2) gene against highly virulent races of Orobanche Cumana Wallr. Theor Appl Genet. 2022;135:501–25.
8. Fernández-Melero B, del Moral L, Todesco M, Rieseberg LH, Owens GL, Carrère S, Chabaud M, Muños S, Velasco L, Pérez-Vich B. Development and characterization of a new sunflower source of resistance to race G of Orobanche Cumana Wallr. Derived from Helianthus anomalus. Theor Appl Genet. 2024;137:56.
9. Su C, Liu H, Wafula EK, Honaas L, de Pamphilis CW, Timko MP. SHR4z, a novel decoy effector from the haustorium of the parasitic weed Striga gesnerioides, suppresses host plant immunity. New Phytol. 2020;226:891–908.
10. Ban X, Qin L, Yan J, Wu J, Li Q, Su X, Hao Y, Hu Q, Kou L, Yan Z, et al. Manipulation of a Strigolactone transporter in tomato confers resistance to the parasitic weed Broomrape. Innov (Camb). 2025;6:100815.
11. Duriez P, Vautrin S, Auriac M-C, Bazerque J, Boniface M-C, Callot C, Carrère S, Cauet S, Chabaud M, Gentou F, et al. A receptor-like kinase enhances sunflower resistance to Orobanche Cumana. Nat Plants. 2019;5:1211–5.
12. Cvejić S, Radanović A, Dedić B, Jocković M, Jocić S, Miladinović D. Genetic and genomic tools in sunflower breeding for Broomrape resistance. Genes (Basel). 2020;11:152.
13. Wang H, Lin YN, Yan S, Hong JP, Tan JR, Chen YQ, Cao YS, Fang W. NRTPredictor: identifying rice root cell state in single-cell RNA-seq via ensemble learning. Plant Methods. 2023;19:119.
14. Yan J, Wang X. Machine learning bridges omics sciences and plant breeding. Trends Plant Sci. 2023;28:199–210.

15. Le Ru A, Ibarcq G, Boniface MC, Baussart A, Muños S, Chabaud M. Image analysis for the automatic phenotyping of Orobanche Cumana tubercles on sunflower roots. Plant Methods. 2021;17:1–14.
16. Xu Y, Zhang X, Li H, Zheng H, Zhang J, Olsen MS, Varshney RK, Prasanna BM, Qian Q. Smart breeding driven by big data, artificial intelligence, and integrated genomic-enviromic prediction. Mol Plant. 2022;15:1664–95.
17. Wang H, Yan S, Wang W, Chen Y, Hong J, He Q, Diao X, Lin Y, Chen Y, Cao Y et al. Cropformer: an interpretable deep learning framework for crop genomic prediction. Plant Commun. 2024;6:101223.
18. Alemu A, Åstrand J, Montesinos-López OA, Isidro y Sánchez J, Fernández-Gónzalez J, Tadesse W, Vetukuri RR, Carlsson AS, Ceplitis A, Crossa J, et al. Genomic selection in plant breeding: key factors shaping two decades of progress. Mol Plant. 2024;17:552–78.
19. Ma X, Wang H, Wu S, Han B, Cui D, Liu J, Zhang Q, Xia X, Song P, Tang C, et al. DeepCCR: large-scale genomics-based deep learning method for improving rice breeding. Plant Biotechnol J. 2024;22:2691–3.
20. Mehta TS, Zakharkin SO, Gadbury GL, Allison DB. Epistemological issues in omics and high-dimensional biology: give the people what they want. Physiol Genom. 2006;28:24–32.
21. Cheng H, Garrick DJ, Fernando RL. Efficient strategies for leave-one-out cross validation for genomic best linear unbiased prediction. J Anim Sci Biotechnol. 2017;8:1–5.
22. Yang Y, Saand MA, Huang L, Abdelaal WB, Zhang J, Wu Y, Li J, Sirohi MH, Wang F. Applications of Multi-Omics technologies for crop improvement. Front Plant Sci. 2021;12:563953.
23. Altman N, Krzywinski M. The curse(s) of dimensionality. Nat Methods. 2018;15:399–400.
24. Yoosefzadeh Najafabadi M, Hesami M, Eskandari M. Machine Learning-Assisted approaches in modernized plant breeding programs. Genes. 2023;14:777.
25. Yu S, Liu L, Wang H, Yan S, Zheng S, Ning J, Luo R, Fu X, Deng X. AtML: an Arabidopsis thaliana root cell identity recognition tool for medicinal ingredient accumulation. Methods. 2024;231:61–9.
26. Sprenger H, Erban A, Seddig S, Rudack K, Thalhammer A, Le MQ, Walther D, Zuther E, Köhl KI, Kopka J, Hincha DK. Metabolite and transcript markers for the prediction of potato drought tolerance. Plant Biotechnol J. 2018;16:939–50.
27. Meng X, Liang Z, Dai X, Zhang Y, Mahboub S, Ngu DW, Roston RL, Schnable JC. Predicting transcriptional responses to cold stress across plant species. Proc Natl Acad Sci U S A. 2021;118:e2026330118.
28. Wang P, Lehti-Shiu MD, Lotreck S, Segura Abá K, Krysan PJ, Shiu SH. Prediction of plant complex traits via integration of multi-omics data. Nat Commun. 2024;15:6856.
29. Wu PY, Stich B, Weisweiler M, Shrestha A, Erban A, Westhoff P, Inghelandt DV. Improvement of prediction ability by integrating multi-omic datasets in barley. BMC Genomics. 2022;23:200.
30. Xu Y, Yang W, Qiu J, Zhou K, Yu G, Zhang Y, Wang X, Jiao Y, Wang X, Hu S, et al. Metabolic marker-assisted genomic prediction improves hybrid breeding. Plant Commun. 2025;6:101199.
31. Azodi CB, Pardo J, VanBuren R, de Los Campos G, Shiu SH. Transcriptome-Based prediction of complex traits in maize. Plant Cell. 2020;32:139–51.
32. Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884–90.
33. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12:357–60.
34. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and samtools. Bioinformatics. 2009;25:2078–9.
35. Liao Y, Smyth GK, Shi W. FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923–30.
36. Robinson MD, McCarthy DJ, Smyth GK. EdgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.
37. Li H, Long C, Xiang J, Liang P, Li X, Zuo Y. Dppa2/4 as a trigger of signaling pathways to promote zygote genome activation by binding to CG-rich region. Brief Bioinform. 2021;22:bbaa342.
38. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:1–13.
39. Love MI, Huber W, Anders S. Moderated Estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
40. Hu B, Zheng L, Long C, Song M, Li T, Yang L, Zuo Y. EmExplorer: a database for exploring time activation of gene expression in mammalian embryos. Open Biology. 2019;9:190054–190054.
41. Chen H, Boutros PC. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. BMC Bioinformatics. 2011;12:35.
42. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25:25–9.
43. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4:44–57.
44. Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, Imamichi T, Chang W. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Res. 2022;50:W216–21.
45. Kanehisa M, Goto S, KEGG. Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30.
46. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L, et al. ClusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innov (Camb). 2021;2:100141.
47. Cawley GC, Talbot NLC. Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers. Pattern Recogn. 2003;36:2585–92.
48. Jiang T, Gradus JL, Rosellini AJ. Supervised machine learning: A brief primer. Behav Ther. 2020;51:675–87.
49. Oikonomou EK, Khera R. Machine learning in precision diabetes care and cardiovascular risk prediction. Cardiovasc Diabetol. 2023;22:259.
50. Pang Z, Xu L, Viau C, Lu Y, Salavati R, Basu N, Xia J. MetaboAnalystR 4.0: a unified LC-MS workflow for global metabolomics. Nat Commun. 2024;15:3675.
51. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.
52. Auriac MC, Griffiths C, Robin-Soriano A, Legendre A, Boniface MC, Munos S, Fournier J, Chabaud M. The penetration of sunflower root tissues by the parasitic plant Orobanche Cumana is intracellular. New Phytol. 2024;241:2326–32.
53. Cellier F, Conéjéro G, Casse F. Dehydrin transcript fluctuations during a day/night cycle in drought-stressed sunflower. J Exp Bot. 2000;51:299–304.
54. Zala HN, Kulkarni KS, Bosamia TC, Shukla YM, Kumar S, Fougat RS, Patel A. Development and validation of EST derived SSR markers with relevance to downy mildew (Sclerospora graminicola Sacc.) resistance in Pearl millet [Pennisetum glaucum (L.) R. Br]. J Plant Biochem Biotechnol. 2017;26:356–65.
55. Blanco-Herrera F, Moreno AA, Tapia R, Reyes F, Araya M, D'Alessio C, Parodi A, Orellana A. The UDP-glucose: glycoprotein glucosyltransferase (UGGT), a key enzyme in ER quality control, plays a significant role in plant growth as well as biotic and abiotic stress in Arabidopsis thaliana. BMC Plant Biol. 2015;15:1–12.
56. Wang H-y, Li P-f, Wang Y, Chi C-y. Jin X-x, Ding G-h. Overexpression of cucumber CYP82D47 enhances resistance to powdery mildew and F.sarium oxysporum F. Sp. cucumerinum. Funct Integr Genom. 2024;24:14.
57. Smith JL, De Moraes CM, Mescher MC. Jasmonate- and salicylate-mediated plant defense responses to insect herbivores, pathogens and parasitic plants. Pest Manag Sci. 2009;65:497–503.
58. Zheng L, Liang P, Long C, Li H, Li H, Liang Y, He X, Xi Q, Xing Y, Zuo Y. EmAtlas: a comprehensive atlas for exploring Spatiotemporal activation in mammalian embryogenesis. Nucleic Acids Res. 2023;51:D924–32.
59. Li H, Long C, Hong Y, Luo L, Zuo Y. Characterizing Cellular Differentiation Potency and Waddington Landscape via Energy Indicator. Research (Wash D C). 2023;6:0118.
60. Kiełbowicz-Matuk A. Involvement of plant C2H2-type zinc finger transcription factors in stress responses. Plant Sci. 2012;185–186:78–85.
61. Han G, Qiao Z, Li Y, Wang C, Wang B. The roles of CCCH Zinc-Finger proteins in plant abiotic stress tolerance. Int J Mol Sci. 2021;22:8327.
62. Zhang H, Gao X, Zhi Y, Li X, Zhang Q, Niu J, Wang J, Zhai H, Zhao N, Li J, et al. A non-tandem CCCH-type zinc-finger protein, IbC3H18, functions as a nuclear transcriptional activator and enhances abiotic stress tolerance in sweet potato. New Phytol. 2019;223:1918–36.
63. Lopez-Bernal D, Balderas D, Ponce P, Molina A. Education 4.0: teaching the basics of KNN, LDA and simple perceptron algorithms for binary classification problems. Future Internet. 2021;13:193.
64. Long C, Li H, Liang P, Chao L, Hong Y, Zhang J, Xi Q, Zuo Y. Deciphering the decisive factors driving fate bifurcations in somatic cell reprogramming. Mol Ther Nucleic Acids. 2023;34:102044.

65. Wasternack C. Action of jasmonates in plant stress responses and development applied aspects. Biotechnol Adv. 2014;32:31–9.
66. Dos Santos CV, Letousey P, Delavault P, Thalouarn P. Defense gene expression analysis of Arabidopsis thaliana parasitized by Orobanche ramosa. Phytopathology. 2003;93:451–7.
67. Albanova IA, Zagorchev LI, Teofanova DR, Odjakova MK, Kutueva LI, Ashapkin VV. Host resistance to parasitic Plants—Current knowledge and future perspectives. Plants. 2023;12:1447.
68. Runyon JB, Mescher MC, Felton GW, De Moraes CM. Parasitism by Cuscuta pentagona sequentially induces JA and SA defence pathways in tomato. Plant Cell Environ. 2010;33:290–303.

Publisher's note