Common genes regulating multiple abiotic stress response in rice: An analysis using supervised and unsupervised machine learning models

Suman Sarkar
ICAR-National Rice Research Institute (ICAR-NRRI)
Parameswaran Chidambaranathan (agriparames07@gmail.com)
ICAR-National Rice Research Institute (ICAR-NRRI)
Kishor Jeughale
ICAR-National Rice Research Institute (ICAR-NRRI)
Bandita Sahoo
ICAR-National Rice Research Institute (ICAR-NRRI)
Raj Kishore Sahoo
ICAR-National Rice Research Institute (ICAR-NRRI)
Hirakjyoti Ray
ICAR-National Rice Research Institute (ICAR-NRRI)
Lopamudra Mohanty
ICAR-National Rice Research Institute (ICAR-NRRI)
Cayalvizhi Balasubramaniasai
ICAR-National Rice Research Institute (ICAR-NRRI)
Sabarinathan Selvaraj
ICAR-National Rice Research Institute (ICAR-NRRI)
Anandan Annamalai
ICAR-National Rice Research Institute (ICAR-NRRI)
Prabhukarthikeyan Seenichamy Rathinam
ICAR-National Rice Research Institute (ICAR-NRRI)
Biswaranjan Behera
ICAR-National Rice Research Institute (ICAR-NRRI)
Awadhesh Kumar
ICAR-National Rice Research Institute (ICAR-NRRI)
Jawahar Lal Katara
ICAR-National Rice Research Institute (ICAR-NRRI)
Devanna Basanvantraya N
ICAR-National Rice Research Institute (ICAR-NRRI)
Lambodar Behera

ICAR-National Rice Research Institute (ICAR-NRRI)
Sanghamitra Samantaray
ICAR-National Rice Research Institute (ICAR-NRRI)
Bhaskar Chandra Patra
ICAR-National Rice Research Institute (ICAR-NRRI)

Research Article

Keywords: Machine learning, Rice, Abiotic stress, Classification problem, ABA, Intrinsic tolerance

Posted Date: July 26th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1861354/v1

Abstract

Rice response to abiotic stresses is generally understood through comparison of gene expression and gene mapping between contrasting tolerant and susceptible genotypes. Machine learning uses large datasets and trains the data through building efficient models for pattern identification and feature importance (gene) prediction. This work explores the potential of using machine learning tools in identifying gene expression models common to different abiotic stress response in rice. For this, 146 rice microarray gene expression samples comprised of drought (57), salinity (35), cold (33) and heat (20) were categorized into two classes namely control, stress and analyzed using supervised (Random forest) and unsupervised (Cluster) machine learning algorithms. The best random forest trained model showed an accuracy of ~ 79% in the classification analysis with an ROC curve area of 0.7963. Besides, genes involved in ABA pathway, flowering, and secondary metabolites (Two PP2Cs, three expressed proteins, Flowering Promoter Factor-like, Terpene Synthase, Gamete EXpressed) were identified as common set of genes for multiple abiotic stress response in rice. Moreover, validation of these eight genes through q-PCR under vegetative stage drought stress identified intrinsic drought tolerance mechanism and drought responsiveness in drought tolerant and rainfed upland, aerobic, and irrigated rice varieties.

Introduction

The advent of sequencing technologies and availability of sequence information in publicly available databases provide valuable resources for fundamental understanding of genes and pathways regulating biological processes. Besides, it provides unique opportunities in comparative analysis of multiple datasets for resolving research problems which would otherwise be less successful in pair-wise datasets comparison. For example, a typical transcriptome analysis provides gene expression information for at least > 10k genes in different samples and time points especially in eukaryotes. Usually expression data is analysed by various statistical models for the identification of differentially expressed genes (Thomas et al, 2001). As compared to individual experiments, analysis of expression data in multiple datasets and experiments is highly challenging but would provide novel gene information and pathways common and characteristics of multiple datasets. Further, it also provides global patterns in gene expression and useful in identification of important features responsible for causal phenotypes. Though, meta-transcriptome analysis are reported (Leimena et al, 2013; Cohen and Leach, 2019), machine learning based computer algorithms used for gene expression analysis is being used increasingly nowadays due to its high prediction power and pattern characterization in the large voluminous data (Pizroonia et al, 2008).

Machine learning uses computational methods and trains the computer for effective data analysis through loss function models (Goodfellow et al, 2016). The process of initial data training for selection of appropriate model parameters and then using the well-trained model for testing the sample input data results in high prediction accuracy and better pattern recognition. Especially, machine learning is nowadays used routinely for solving the genetics-based research problems (Libbrecht and Noble, 2015). In machine learning, there are unsupervised and supervised machine learning models employed in the

analysis of genetic data sets which includes sequence information of genes and promoters, expression data of genes, and epigenetic datasets (Shipp et al, 2002).

Machine learning task are generally categorized into supervised and unsupervised learning methods (Lopez et al, 2018). The unsupervised machine learning methods are useful in finding the patterns present in the genetic (expression) information. However, the data are not labelled in unsupervised machine learning tools. In contrast, supervised machine learning comprises of labelled data used for training the models through computational methods and algorithms (Abdulquader et al, 2020). Further, trained models are used for solving different types of research problems broadly categorized into classification and regression analysis (Liaw and Weiner, 2002). Besides, machine learning algorithms are capable of finding hierarchical order of feature importance, prediction of complex traits, analysis of population genetic drift, location of causal genes for complex diseases and traits (Libbrecht and Noble, 2015).

Rice is one of the major food crops which are genetically and agronomically adapted to high water requirements (~ 2500–5000 litres of water for 1 kg of paddy) for producing a unit quantity of economic product. Thus, evaluation and improvement of rice varieties in limited/reduced irrigation strategiesfor high genetic gain is becoming an important research area recently because of anticipated effects of climate change (Tuong and Bouman, 2003). Generally it is presumed that multiple abiotic stress tolerance in rice varieties are associated with greater adaptation to growth in unfavourable ecologies such as limited water availability or rainfed cultivation. Previously several major genes, quantitative trait loci, genes under linkage disequilibrium associated with abiotic stress tolerance were identified for regulating drought tolerance in rice (Bernier et al, 2007). However, though major genes and pathways are well characterized for individual abiotic stresses in rice namely drought and salinity, overall common mechanism regulating abiotic stress tolerance needs to be understood with higher degree of clarity for finding novel solutions to address the requirement of higher yield under rainfed drought prone conditions. In the present work, gene expression data of multiple abiotic stresses were analyzed using machine learning models for the identification of genes and pathways capable of classifying the stress and non-stress (control) conditions in rice.

Results

2.1 Number of microarray samples used in analysis

GEO database was used for the microarray data retrieval. In total, mean expression value of 146microarray samples (57 drought, 20 heat, 33 cold, and 35 salinity experiments) were taken for the analysis. Of these samples, 54 samples in the retrieved data were labelled as control samples and 92 as stress treatment samples and used in machine learning analysis for the identification of causal genes involved in abiotic stress response. The list of samples taken for the analysis is given in SupplementaryFig. 2.Each data samples taken for the analysis was comprised of 57,194 set of genes/probes and total numbers of data normalized expression data analyzed was ~ 8,350,324.

2.2 K means Clustering

The samples were normalized using min-max and quantile normalization method. Further, K means analysis with six clusters for the 146 samples showed maximum variation of ~ 83% between the clusters. The grouping of the probes/genes into six clusters is given in Fig. 1.Similarly, clara medoids with pamlike algorithm was used to retrieve a set of thousand probes/genes representing the six different clusters.

2.3 Gene ontology analysis

The selected 1000 probes/genes representing six clusters were initially analyzed for their gene function using gene ontology analysis. The analysis for biological process indicated that the genes involved in cellular signaling, signal transduction, and regulation of biological process were highly predominant in the biological process (Fig. 2). Similar analysis for the metabolic process showed amino acid metabolism, protein, and nitrogen metabolism were highly enriched. Further, among the cellular process, organelle specific, membrane related, and mitochondria genes were highly enriched in the selected genes representing all the six clusters (Fig. 2).

2.4 Boruta and feature importance using random forest

Boruta wrapper algorithm for random forest analysis was used for the identification of the important genes capable of classifying control and stress samples. A five hundred replication of the thousand selected genes in Boruta identified eight genes as important, 983 were found to be unimportant and ten as tentative attributes. Further, selected eight genes were used as factors in the random forest analysis(Table 1). Initially, random forest algorithm was used in all the 146 samples which comprised of 54 control(non-stress) and 92 stress samples. The random forest model with 500 numbers of trees and four variables at each split identified OOB error rate of 21.92%. Besides, among the class error, 20 control samples were classified as stress samples with an error rate of 33% and 15 stress samples were classified as control samples with an error rate of 15.21%. Further, RF plot for the number of trees also showed OOB error rate in the range of 23–24% in the analysis. The specificity test of the model with area under the ROC was found to be 0.8369 indicating high specificity of the random forest model.

The samples were also divided into training and testing datasets with 30% samples were grouped as test data for random forest analysis. Fine tuning of tree parameters with 1–8 variables at each split and number of trees between 100 and 500 for training dataset identified the least OOB error rate of 27% for 400 numbers of trees with one variable at each split. Further, test data analysis using this optimized model showed an error rate of 19.57% with class error of only 9% for stress samples(Fig. 3). However, class error of control samples was found to be very high (46%), i.e. the model classified six control samples as stress samples. Besides, ROC curve with area under the plot was identified to be 0.7963 indicating high sensitivity for the trained random forest model. Further, random forest model for all the data was used for the variable important analysis. This analysis showed high mean decrease in variable impurity was observed for both the PP2C genes indicating its significant importance(Supplementary Fig. 3). Additionally, since the model showed high class error for control samples, it was presumed that

most of the identified important genes might show expression also into non stress (control) samples. Thus, box plot analysis of the important genes was performed and it showed most of the selected genes were expressed in both the conditions except a PP2C and TPS gene. Besides, most of the genes were up regulated in stress conditions except an expressed protein. Also, high degree of up regulation was observed for a PP2C and two expressed protein(Supplementary Fig. 4).

2.5 Drought stress analysis

After initiation of stress treatments, leaf rolling was initially observed in the varieties to understand its response to drought stress. Before stress, all the genotypes showed a similar leaf rolling score of zero. Further, symptoms of drought stress were observed within three days of treatment as few varieties namely CR Dhan 201, Vandana, and Naveen showed leaf rolling symptoms. After eight days of stress treatment, mean leaf rolling score was 7 or more for all the varieties except Way Rarem (Score:5). In addition to leaf rolling, other morphological and physiological parameters were also analyzed. On an average, mean plant height for all the varieties reduced by 8.76 cm in drought stressed plants. Besides, a least reduction in plant height was observed in CRDhan201 (3.22 cm) and the highest reduction in Way Rarem (14.02 cm). Similarly, Way Rarem also showed higher reduction in tiller number (1.6) as compared to other varieties. However, a least reduction in tiller number was observed in Naveen variety (0.24). Also, mean leaf length of all the varieties in stress treatments was reduced by an average value of 3.3 cm(Fig. 4).

Chlorophyll status of the varieties was measured using the SPAD meter in both control and stress treatments. All the varieties showed reduction in the SPAD value. However, as compared to control, least reduction in SPAD value was observed in Way Rarem (7.32%) variety. All other varieties, namely Vandana (17.46%), Sahbhagi Dhan (20.10%), Naveen (28%), and CR Dhan 201(35.35%) showed higher reduction in SPAD in stress conditions. Similarly, Way Rarem (20.68%) showed the least reduction in RWC under stress treatments as compared to Sahbhagi Dhan (22.78%), Naveen (27.2%), CR Dhan 201 (32.1%), and Vandana (55.45%).

Principal component analysis was performed for all the phenotypic traits to understand the grouping of varieties with traits. Principal component 1 (PC1) showed maximum variation of 53.83% and differentiated varieties, CR Dhan 201, Vandana in one component and Way Rarem, Naveen, and Sahbhagi Dhan in the opposite component. Similarly, PC2 showed 29.76% variation and grouped only Way Rarem and Vandana in a component. Further, biplot analysis for traits and varieties showed RWC and SPAD of both control and stress were grouped with Way Rarem variety in PC1 along with Sahbhagi Dhan (drought tolerant) and Naveen. In contrast, no. of tillers, plant height, and leaf length were grouped with CR Dhan 201 and Vandana. Additionally, cumulative variation observed in both the components was ~ 83%(Fig. 5).

2.6 Quantitative PCR analysis

The expression of eight genes analyzed through qPCR is given in Fig. 6. It showed seven genes were up regulated in drought stress condition relative to control exceptan expressed protein (LOC_Os10g35340). This gene was invariably down regulated in all genotypes under drought stress. Further, higher magnitude

of up regulation under stress was observed in varieties namely CR Dhan 201, Vandana, and Naveen. Additionally, comparison between the genotypes in control and stress conditions in relation to Naveen variety also showed major differences in the expression pattern of genes between varieties. For example,relative fold change expression of six genes (Two PP2Cs, two expressed protein, FPF1, and GEX) in Way Rarem was higher than Naveen under control condition. Similar pattern was also observed in drought tolerant Sahbhagi Dhan though the fold change difference was less relative to Naveen.In drought stress, Way Rarem showed similar expression to that of Naveen expect two genes (PP2C: LOC_Os05g04360 and Expressed protein: LOC_Os02g36200) were down regulated. Besides, expression level in Sahbhagi Dhan was also similar to Naveen under drought stress. In Vandana, two genes (Expressed protein: LOC_Os02g36200 and GEX:LOC_Os09g27040) were up regulated under stress conditions in relation to Naveen(Fig. 7a).Further,principal component analysis was used to group the varieties based on the expression pattern in control and stress condition. In control condition, first principal component explained a maximum of 95.08% variation and second component differentiated Way Rarem, Sahbhagi Dhan, and CR Dhan 201 to that of Naveen and Vandana varieties. Biplot analysis for gene expression and varieties showed PP2Cs, FPF1, and an expressed protein (LOC_Os05g43860) was grouped along with Way Rarem and Sahbhagi Dhan. Though principal components under stress conditions also grouped Way Rarem and Sahbhagi Dhan in one component, only three genes namely both the PP2Cs and an expressed protein (LOC_Os02g36200) were grouped along with the two varieties. Additionally, Vandana was also grouped together with Sahbhagi Dhan and Way Raremin PCA analysis under stress conditions (Fig. 7b).

Discussion

In this work, normalised expression data of genes in multiple abiotic stress conditions in rice was analyzed using machine learning algorithms for the identification of candidate genes associated with abiotic stress tolerance in rice. Initially, unsupervised learning algorithm predicted the pattern present in the gene expression data and then boruta and random forest feature selection and importance analysis identified genes which classified the control (non-stress) and stress samples with ~ 79% accuracy(Fig. 8).The salient findings of this study is given below,

3.1 Constitutive expression of PP2Cs of ABA signaling as major component of multiple abiotic stress response

ABA pathway is well recognized for abiotic stress tolerance in plants (Raghavendra et al, 2010). ABA signalling components includes ABA receptor (PYR/PYL/RCAR), co-receptor (PP2Cs), kinase (SNF-1), and downstream transcription factors (ABI/ABF) regulated through ABA concentrations, transcriptional, and post transcriptional modifications in abiotic stress response (Ali et al, 2020). Our findings identified that two of the ABA co-receptor (PP2Cs) multigene family are major important genes for multiple abiotic stress tolerance in rice. Previously, PP2Cs are reported to modulate drought stress response (Santiago et al, 2009) and this gene family evolved in land plants as intrinsic regulators of desiccation tolerance (Komatsu et al, 2013). Besides, highly expressed PP2Cs was reported to regulate drought resistance in Arabidopsis (Bhaskara et al, 2012). The present study identified two PP2Cs (LOC_Os06g48300:*OsPP91*;

LOC_Os05g04360:*OsPP72*) which were highly up regulated in varietiesnamely CR Dhan 201, Naveen, Sahbhagi Dhan, and Vandana under stress conditions and constitutively highly expressed in Way Rarem variety indicating its possible function in drought tolerance signaling. In support of our observation, *OsPP72* and *OsPP91* belonging to PP2Cs subfamily G and F, respectively was found to be highly expressed in salinity and drought stress in rice (Singh et al, 2010). Besides, estradiol induced transient expression of *OsPP91* before stress (priming) enhanced the osmotic and drought tolerance during stress (Yu et al, 2018) indicating constitutive high expression of *OsPP91* might be one of the major tolerance mechanism in rice for drought stress tolerance. Therefore, Way Raremvariety could possess innate tolerance mediated by *OsPP91* high expression in non stress condition in addition to stress responsive QTLs such as qDTY12.1 (Mishra et al, 2013). Furthermore, *OsPP91* was also reported to co-localize with genomic hotspots regions for salinity tolerance in rice meta-QTL analysis (Mirdar et al, 2020) and cold tolerance (Najeeb et al, 2020). Probably, as reported by Yu et al (2018), suberization of root cells mediated by over expression of *OsPP91* could conserve water in roots and enhances tolerance to multiple abiotic stress tolerance. Thus, multi-functional abiotic stress response of *OsPP91* could be exploited in rice improvement programs.

3.2 Terpene synthesis for abiotic stress response

Two varieties namely WayRarem and Sahbhagi Dhan showed up regulation of terpene synthase (LOC_Os04g27670: TPS) to the tune of more than 20 fold as compared to other varieties in our analysis. Terpenes are important in interaction of plants with biotic and abiotic stress stimuli in the environment (Falara et al, 2011). For example, forest trees respond to drought stress by regulating the synthesis of terpenes (Llusia et al, 2016). In maize, accumulation of terpenoids was ABA inducible, associated with drought tolerance, and mutants of terpene synthesis were highly sensitive (Vaughan et al, 2015). Moreover, previous reports in rice identified terpene synthase gene (LOC_Os04g27670) for brown plant hopper tolerance (Kamolsukyunyong et al, 2013). Thus, potential antioxidant role of terpene related compounds in abiotic stress in rice needs to be further explored. Further, it is reported that terpene synthesis during drought stress assist in the synthesis of other major terpenoids in plants namely ABA, chlorophyll, and carotenoids indicating crosstalk between terpene synthesis and ABA signalling (Jogawat et al, 2021).

3.3 Novel expressed proteins for drought stress response

The present study has identifiedthree expressed genes (LOC_Os10g35340; LOC_Os05g43860; LOC_Os02g36200) showing high importance for drought stress response. All these genes are not functionally annotated and further characterization is required for its functional relationship with abiotic stress tolerance. Specifically, one of the expressed proteins (LOC_Os10g35340) was found to be highly down regulated in all the varieties indicating that down regulation might be related with drought response. The loss-of-function allele of these genes might provide better insights for drought tolerance.

3.4 Differential response of rice varieties under drought stress conditions

Phenotypic and expression analysis indicated that the response of rice varieties may be governed by the intrinsic drought tolerance mechanisms and also drought responsiveness. For example, Way Rarem showed higher expression of PP2Cs even under control condition and maintained higher RWC under stress conditions reflecting better intrinsic drought tolerance and responsiveness under stress. In support of our observation, constitutive expression of PP2C of rice in Arabidopsis showed improved abiotic stress tolerance (Singh et al, 2015). Though CR Dhan 201, Vandana, and Naveen showed induced PP2Cs expression, comparatively constitutive high expression may provide higher drought tolerance. Further, a previous report in pearl millet identified mechanism related to conservation of water in plants was related with terminal drought tolerance (Kholova et al, 2010). In support of this observation, suberized layers in roots of *OsPP91*over expressed line was reported to reduce water loss (Yu et al, 2018), thus it is plausible that constitutive high expression of *OsPP91* and *OsPP72*in Way Rarem could contribute to intrinsic drought tolerance through one mechanism of preventing water loss within root cells. This notion requires systemic investigation for confirmation. Further corroborating with intrinsic drought tolerance mechanism, stress imposed in this experiment was very rapid and within 2–3 days, few varieties started leaf rolling symptoms indicating a strong intrinsic drought tolerance mechanism would be essential for survival under rapid stress induction. Besides, it was observed in Way Rarem which showed relatively better tolerance and delayed leaf rolling symptoms indicating unique mechanism for intrinsic drought tolerance. In support, constitutive high expression of salinity related genes was reported for high degree of salinity tolerance in Pokkali cultivar of rice,in addition to potassium and sodium uptake dynamicsduring salt stress (Lakra et al, 2019). The varieties taken in our analysis are highly popular in water limited ecologies such as upland, rainfed, and aerobic conditions. Thus, intrinsic stress tolerance and stress responsiveness might be an important factor for the popularity of these varieties in unfavorable ecologies. The mechanistic insights in to intrinsic drought tolerance could be highly useful not only for rainfed cultivation but also for the yield improvement for reduced irrigation strategies.

3.5 Utility of machine learning in understanding multiple abiotic stress response

The identified eight genes could classify the control and stress samples up to 79% accuracy. Besides, error rate observed was 21%. This indicated that the selected gene expression pattern in multiple abiotic stresses would be functionally significant. In agreement with our observation, potential of machine learning analysis using genes, promoter sequences, and expression datafor gene function prediction was highlighted by Mahmood et al, (2020). Further, classification accuracy of 79% in this analysis is comparable to the prediction of sub cellular localization (85%), protein-protein interaction (80%), and phenotype prediction (87%) of genes using machine learning approaches (Mahmood et al, 2020). However, higher classification accuracy between 85% and 100% using feature selection algorithms were also reported using microarray datasets of different cancer and other tissues in humans (Pirooznia et al, 2008; Cui et al, 2019). Though, in these studies at least 15 to 20 important genes were used for classification after feature selection as compared to only eight genes used in our analysis indicating improvement in feature selection models could greatly improve the prediction accuracy. Besides, similar analysis using microarray data for prediction of abiotic stress responsive genes in Arabidopsis showed

prediction accuracy of 67 to 84%, even after using few thousands of informative genes (Ma et al, 2014). Additionally, prediction of nitrogen responsive genes in Maize and Arabidopsis using machine learning also used only 15 informative genes and showed overall prediction accuracy ranged from 59 to 79% and varied among genotypes between 5% and 99% (Cheng et al, 2021). Therefore, plausible explanation for prediction accuracy of 79% in our analysis would be though microarray analysis are generally performed between control and stress conditions, tolerant genotypes taken in the study most probably exhibits stress responsive gene expression even under control conditions characterized as intrinsic abiotic stress tolerance. This mechanism might interfere with classification analysis using machine learning models for the prediction of important genes related to stress response. In spite of genotype and sample biasness, usage of deep learning algorithms and improved feature selection strategies could significantly improve the prediction accuracy (Wang et al, 2020).

Conclusion

Machine learning has great potential for the identification of important features in the complex gene expression data. Our analysis identified eight genes which could predict different abiotic stress experimental samples distinctly into control and stress classes with high accuracy of ~ 79%. Further, identified genes validated through qPCR in rice varieties of rainfed, upland, aerobic, and irrigated ecologies under drought stress showed constitutive high expression is associated with relatively higher tolerance indicative of intrinsic drought tolerance mechanism. Additionally, three novel expressed proteins were identified which needs further characterization for understanding its functional importance under drought stress. However, further improvement in prediction accuracy using different normalization methods and important variable selection are required for comprehensive understanding of multiple abiotic stress response in rice using machine learning models.

Methodology

4.1Data Retrieval

The expression data of genes under different abiotic stress samples (Drought, heat, salinity, and cold) in rice using affymetrix rice genome array (Platform: GPL2025) was downloaded from the gene expression omnibus database (GEO) of NCBI (https://www.ncbi.nlm.nih.gov/geo). The list of accession number for the experiments retrieved from GEO database is given in Supplementary Table 1. This data comprised of expression profiling of genes in the array for the samples of leaf, roots, shoots, and young panicles at three developmental stages (rooting, tillering and panicle stage). Mostly, the analyzed array experiments were direct comparison between abiotic stress tolerant and susceptible genotypes of rice performed under different intensity of stress and also varied in stages of treatments. In total, retrieved dataset consists of 146 samples of control and stress treatments with each sample comprised of 57,194 RMA normalized signal values of genes/probes.

4.2 Data labeling

R Version 3.6.0 (https://www.R-project.org/.)was used for all the downstream analysis of the retrieved expression profiling datasets. The list of packages used in our analysis is given in Supplementary Table 2.The retrieved GEO datasets comprised of samples with three biological replicates. Therefore, mean of RMA normalized genes/probes values in three biological replicates was calculated initially and further used in the analysis. Then, samples were labelled as control (non stress) and stress classes as per the accession id and title of the replicated samples given in the microarray data table of each experiment in the GEO database. For example, in experiments with different time points (0, 1hr, 3hr, and 6 hr), 0hr was labelled as control sample and remaining three time points were labelled as stress samples. Accordingly, multiple abiotic stress (drought, heat, salinity, and cold) and control samples (Supplementary Table 1)were relabelled as only two different classes namely 'control' and 'stress' and used in the analysis of classification problem through machine learning models. This pattern of labelling was assumed to be useful in identification of common abiotic stress responsive genes and pathways in rice.

4.3 Data normalization

The expression datasets comprised of normalized expression values of multiple independent experiments performed by different researchers in independent labs and variation within and between the experiments would interfere with data analysis and feature selection (gene selection) (Thompson et al, 2016). Thus, data normalization was performed to reduce the variation present in the retrieved microarray datasets. Firstly, min-max normalization which conserves the pattern within the data and also converts the expression values between 0 and 1 was applied as reported previously by Patro and Sahu (2015), Kappal (2019), Gokhan et al, (2019) and Henderi et al (2021). After min-max normalization, varied distribution in expression intensity of probes/genes present between the independent experimental samples were normalized using quantile normalization method as reported by Gallon et al, (2013),Qiu et al, (2013), and Liu et al, (2019).

4.4 Data clustering

Clustering of gene expression data result in grouping of individual probes/genes based on similar expression values (Jiang et al, 2004).In clustering, highly expressed, medium, and low expressed genes would be partitioned into separate clusters (Chandrasekhar et al, 2013; Oyelade et al, 2016; Saadeh et al, 2020). Therefore, K-means and CLARA medoids (Andreopoulos et al, 2009) clustering approaches which clusters the data using partitioning around medoids (PAM) algorithm was used to group the microarray gene expression data in to different clusters (Likas et al, 2003). Briefly, parameters used were: number of clusters = 2–6, iter.max = 10, and nstart = 100 for K means, and six clusters, number of samples = 146, sample size = 1000 along with pamlike algorithm in clara medoids.

4.5 Feature selection and importance analysis

Feature selection (gene selection) is an important step which identifies variables capable of classifying the samples classes into control (non-stress) and stress conditions. In microarray datasets, gene

expression data of several thousands ofgenes are to be analyzed to identify the few important genes having expression pattern to differentiate control and stress classes ofsamples. For this, Boruta algorithm (Kursa and Rudnicki, 2010), a wrapper algorithm used along with the random forest classifier (Qi, 2012) was used for selection of important genes. In Boruta method, deviation in mean loss in accuracy of tress built using variables (genes) and computed as Z score was used for the identification of important variables capable of classifying control and stress samples (Degenhardt et al, 2019; Ab Aziz et al, 2020). The maxrun parameter of 500 in Boruta model was used for the differentiation of variables (genes/probes) into important, unimportant, and not considered variables. Further, only the important variables (genes/probes) identified using the Boruta algorithm was further used in supervised ensemble based random forest package (Liaw and Wiener, 2002). Random forest classifier was performed for two types of data partition. In one, all the data comprised of 146 samples had been used for the random forest classification analysis. In another, 70% of data was used for the training of the random forest model, and remaining 30% was used for testing the optimized trained model. The model tuning in random forest was performed using the following parameters: ntree = 100 to 500, and mtry = 1–8. The out of box (OOB) error was measured and the model showing least OOB error in the training data was used for the OOB prediction in test samples. Further, receiver operating characteristic curve (ROC) curve was used for determination of area under curve for the optimized random forest model as reported by Naghibi et al, (2017). Then, variance important plot was used for determining the sequence of importance based on mean decrease in variable impurity.

4.6 Drought stress treatments

An experiment was conducted in the green house facility at the National Rice Research Institute (ICAR-NRRI), Cuttack, Odisha, India (20°27'9"N, 85°56'25"E), during Aug to Dec month of 2021. Five popular varieties namely CR Dhan 201, an aerobic variety, Way Rarem, an upland variety, Naveen, an irrigated/boro variety, Sahabhagi Dhan, a drought tolerant variety, and Vandana, a rainfed upland variety (Pathak et al, 2019; Basu et al, 2017; Mishra et al, 2013) were used for seedling stage drought stress experiment. The seeds (50 nos) of the five varietieswere placed on the tissue paper kept in the petriplate and ~ 5–10 mL of distilled water was added to moisten the tissue paper and petriplates were kept in dark condition under room temperature for germination. Each plates were watered regularly to moisten the tissue paper during the germination and post germination stages. After 10 days of sowing, seedlings were transplanted into 6 inch pots filled with wetland soil for seedling establishment. For each variety, five pots were maintained for control and stress treatments and 2–3 seedlings were transplanted and irrigated to full saturation (~ 3–5 cm of standing water). Further, 25 days after sowing(25 DAS) at 3–4 leaf stage of seedlings, stress treatment was imposed by withholding the irrigation in one set of pots maintained for stress treatments. The control pots were irrigated daily and water level was maintained up to ~ 3 to 5 cm from surface soil. Further, stressed plants were observed for leaf rolling under drought stress as given by Standard evaluation system (SES) in rice (IRRI, 2002) and when the leaf rolling was observed in all the varieties, shoot samples were collected, immediately dipped in liquid nitrogen, and kept

at -80°C for RNA expression analysis. For each variety, three biological replicates for control and stress shoot samples were collected and stored as mentioned above(Supplementary Fig. 1).

4.7 Phenotypic Analysis

Leaf fresh weight (weight of 3rd leaf samples in control and stress condition), and dry weight (weight of 3rd leaf samples in control and stress condition after drying in hot air oven @70°C for five days) were measured by collecting five leaf samples for each variety under both control and stress conditions and analysis was done as reported by Farooq et al, (2010). Besides, leaf samples (~ 0.1g) were also collected for the relative water content measurement (RWC) in control and stress samples as reported by Farooq et al, 2010. Further, morphological parameters such as plant height (height of seedlings from the base of soil), leaf length (length of 3rd completely expanded leaves), tiller number (no. of tillers) and chlorophyll index (chlorophyll readings measured using SPAD meter in the fully expanded third leaf) were also collected for the analysis as reported by Mishra and Panda, (2017).The morphological data, RWC, and SPAD measurements were taken from five seedlings in different pots from both control and stress treatments of each varieties.

4.8 Quantitative PCR analysis

Total RNA was isolated using RNeasy kit, Quigen, Germnay as per protocol mentioned in the kit. Quality of the isolated total RNA was checked through running in 1.2% agarose gel electrophoresis and quantity was determined using nanodrop spectrophotometer. DNA contamination in the isolated RNA was removed using DNase I from NEB, USA. Further, 1.5 µg of DNase I treated total RNA was used for first strand cDNA synthesis using qScript cDNA synthesis kit, Quantabio, USA. Then, gene expression was studied using 1µL of cDNA and SYBR premix Ex Taq, Takara, Japan in a 10 µL reaction volume consisting of 5 µL of SYBR buffer (2x), 1 µL of primer (0.05 µM), and 3µL of double autoclaved water. The gene expression was quantified using three biological and three technical replicates. CFX96-Touch realtime PCR system, Biorad, USA was used for gene quantification with following cycles, initial denaturation of 94°C for 2 min, followed by 40 cycles of 94°C for 15s (denaturation), 60°C for 30s (annealing),72°C for 30s (extension) and followed by melt curve analysis. The uniqueness of the amplicon was analyzed using the melt curve and compared with non templatecontrol of the respective genes. Rice 25s gene was used as internal reference gene for normalization. The relative change in fold difference was calculated by delta deltact method as previously reported by Livak and Schmittgen, (2010). The primers used in the analysis are given in Supplementary Table 3.

Declarations

This manuscript is not submitted elsewhere for review or publication

Competing interests

Authors declare no financial interests or any other competing interests.

Authors contribution

PC- Conceptualization, data analysis, manuscript writing and finalization. SS, KJ, BS- Data retrieval, ML analysis, manuscript primary draft writing. RKS, HR, LM, CB, SS- Statistical analysis, gene expression analysis, manuscript finalization. PSR, AK, DB, and JLK-Model validation, primer designing and validation, and manuscript editing. AA, BB: Drought stress experiment management. LB, SS, and BCP - Manuscript editing, overall coordination, and manuscript finalization.

Funding

This work is supported by general institute grant from ICAR, New Delhi

Acknowledgements

We sincerely acknowledge the Director General, Indian Council of Agricultural Research (ICAR), New Delhi for the institute financial support and the Director, National Rice Research Institute (ICAR-NRRI), Cuttack for providing lab and other facilities for performing the experiments.

References

1. Ab Aziz, N. A., Besar, R., Mohd Ali, N., 2020. Comparison of microarray breast cancer classification using support vector machine and logistic regression with LASSO and boruta feature selection. Indonesian J. Electrical Eng. Comput. Sci. 20(2), 712-719.
2. Andreopoulos, B., An, A., Wang, X., Schroeder, M., 2009. A roadmap of clustering algorithms: finding a match for a biomedical application. Briefings Bioinform. 10(3), 297-314.
3. Basu, S., Jongerden, J., Ruivenkamp, G., 2017. Development of the drought tolerant varietySahbhagi Dhan: exploring the concepts commons and community building. Int.J. Commons. 11(1).
4. Bolstad, B. M., Bolstad, M. B. M., 2013. Package 'preprocessCore'.
5. Chandrasekhar, T., Thangavel, K., Elayaraja, E., Sathishkumar, E. N., 2013, March. Unsupervised gene expression data using enhanced clustering method. In 2013 IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology(ICECCN),518-522.
6. Cheng, C. Y., Li, Y., Varala, K., Bubert, J., Huang, J., Kim, G. J., Coruzzi, G. M., 2021. Evolutionarily informed machine learning enhances the power of predictive gene-to-phenotype relationships. Nat.Commun. 12(1), 1-15.
7. Cui, S., Wu, Q., West, J., Bai, J., 2019. Machine learning-based microarray analyses indicate low-expression genes might collectively influence PAH disease. PLoS Comput. Biol. 15(8), e1007264.
8. Degenhardt, F., Seifert, S., Szymczak, S., 2019. Evaluation of variable selection methods for random forests and omics data sets. Briefings in Bioinform. 20(2), 492-503.

9. Farooq, M., Kobayashi, N., Ito, O., Wahid, A., Serraj, R., 2010. Broader leaves result in better performance of indica rice under drought stress. J. Plant Physiol. 167(13), 1066-1075.
10. Gallón, S., Loubes, J. M., Maza, E., 2013. Statistical properties of the quantile normalization method for density curve alignment. Math.Biosci. 242(2), 129-142.
11. Gökhan, A. K. S. U., Güzeller, C. O., Eser, M. T., 2019. The effect of the normalization method used in different sample sizes on the success of artificial neural network model. International J. Assess. Tools Edu. 6(2), 170-192.
12. Hauser, J. R., Zettelmeyer, F., 1997. Metrics to Evaluate R, DE. Res. Technol. Manag. 40(4), 32-38.
13. Henderi, H., Wahyuningsih, T., Rahwanto, E., 2021. Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer. Int. J. Inform. Syst. 4(1), 13-20.
14. Jiang, D., Tang, C., Zhang, A., 2004. Cluster analysis for gene expression data: a survey. IEEE Trans.Knowl.Data Eng. 16(11), 1370-1386.
15. Kamolsukyunyong, W., Sukhaket, W., Ruanjaichon, V., Toojinda, T., Vanavichit, A., 2013. Single-feature polymorphism mapping of isogenic rice lines identifies the influence of terpene synthase on brown planthopper feeding preferences. Rice 6(1), 1-9.
16. Kappal, S., 2019. Data normalization using median median absolute deviation MMAD based Z-score for robust predictions vs. min–max normalization. Lond J. Res. Sci. Nat. Formal.
17. Kholová, J., Hash, C. T., Kakkera, A., Kočová, M., Vadez, V., 2010. Constitutive water-conserving mechanisms are correlated with the terminal drought tolerance of pearl millet [Pennisetum glaucum (L.) R. Br.]. J. Exp. Bot. 61(2), 369-377.
18. Kursa, M. B., Rudnicki, W. R., 2010. Feature selection with the Boruta package. J. Stat. Softw. 36(11), 1-13.
19. Lakra, N., Kaur, C., Singla-Pareek, S. L., Pareek, A., 2019. Mapping the 'early salinity response'triggered proteome adaptation in contrasting rice genotypes using iTRAQ approach. Rice 12(1), 1-22.
20. Lemenkova, P., 2019. K-means Clustering in R Libraries {cluster} and {factoextra} for Grouping Oceanographic Data. Int. J. Appl. Math. 2(1), 1-26.
21. Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R news 2(3), 18-22.
22. Likas, A., Vlassis, N., Verbeek, J. J., 2003. The global k-means clustering algorithm. Pattern Recognit. 36(2), 451-461.
23. Liu, X., Li, N., Liu, S., Wang, J., Zhang, N., Zheng, X., Cheng, L., 2019. Normalization methods for the analysis of unbalanced transcriptome data: a review. Front. Bioeng. Biotechnol. *7*, 358.
24. Livak, K. J., Schmittgen, T. D., 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2− ΔΔCT method. Methods 25(4), 402-408.
25. Ma, C., Xin, M., Feldmann, K. A., Wang, X., 2014. Machine learning–based differential network analysis: a study of stress-responsive transcriptomes in Arabidopsis.Plant Cell26(2), 520-537.

26. Mahood, E. H., Kruse, L. H., Moghe, G. D., 2020. Machine learning: A powerful tool for gene function prediction in plants. Appl. Plant Sci. 8(7), e11376.
27. Mirdar Mansuri, R., Shobbar, Z. S., BabaeianJelodar, N., Ghaffari, M., Mohammadi, S. M., Daryani, P., 2020. Salt tolerance involved candidate genes in rice: an integrative meta-analysis approach. BMC Plant Biol. 20(1), 1-14.
28. Mishra, K. K., Vikram, P., Yadaw, R. B., Swamy, B. P., Dixit, S., Cruz, M. T. S., Kumar, A., 2013. qDTY12. 1: a locus with a consistent effect on grain yield under drought in rice. BMC Genet. 14(1), 1-10.
29. Mishra, S. S., Panda, D., 2017. Leaf traits and antioxidant defense for drought tolerance during early growth stage in some popular traditional rice landraces from Koraput, India. Rice Sci. 24(4), 207-217.
30. Naghibi, S. A., Ahmadi, K., Daneshi, A., 2017. Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping. Water Resour. Manag. 31(9), 2761-2775.
31. Najeeb, S., Ali, J., Mahender, A., Pang, Y. L., Zilhas, J., Murugaiyan, V., Li, Z., 2020. Identification of main-effect quantitative trait loci (QTLs) for low-temperature stress tolerance germination-and early seedling vigor-related traits in rice (Oryza sativa L.). Mol. Breed. 40(1), 1-25.
32. Oyelade, J., Isewon, I., Oladipupo, F., Aromolaran, O., Uwoghiren, E., Ameh, F., Adebiyi, E., 2016. Clustering algorithms: their application to gene expression data. Bioinform. Biol. Insights. 10, BBI-S38316.
33. Pathak, H., Parameswaran, C., Subudhi, H. N., Prabhukarthikeyan, S. R., Pradhan, S. K., Anandan, A., Sarkar, R. K., 2019. Rice Varieties of NRRI: Yield, Quality, Special Traits and Tolerance to Biotic andAbiotic Stresses.
34. Patro, S.,Sahu, K. K., 2015. Normalization: A preprocessing stage. arXiv preprint arXiv:1503.06462.
35. Pirooznia, M., Yang, J. Y., Yang, M. Q., Deng, Y. (2008). A comparative study of different machine learning methods on microarray gene expression data. BMC Genet. 9(1), 1-13.
36. Qi, Y., 2012. Random forest for bioinformatics. In Ensemble machine learning. Springer, Boston, MA, 307-323.
37. Qiu, X., Wu, H., Hu, R., 2013. The impact of quantile and rank normalization procedures on the testing power of gene differential expression analysis. BMC Bioinform. 14(1), 1-10.
38. R Core Team., 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URLhttps://www.R-project.org/.
39. Rieder, V., Schork, K. U., Kerschke, L., Blank-Landeshammer, B., Sickmann, A., Rahnenführer, J., 2017. Comparison and evaluation of clustering algorithms for tandem mass spectra. J.Proteome Res. 16(11), 4035-4044.
40. Saadeh, H., Al Fayez, R. Q., Elshqeirat, B., 2020. Application of K-Means Clustering to Identify Similar Gene Expression Patterns during Erythroid Development. Int. J. Mach. Learn. Comput. 10.
41. Singh, A., Giri, J., Kapoor, S., Tyagi, A. K.,Pandey, G. K., 2010. Protein phosphatase complement in rice: genome-wide identification and transcriptional analysis under abiotic stress conditions and

reproductive development. BMC Genet. 11(1), 1-18.

42. Singh, A., Jha, S. K., Bagri, J., Pandey, G. K., 2015. ABA inducible rice protein phosphatase 2C confers ABA insensitivity and abiotic stress tolerance in Arabidopsis. PloS One 10(4), e0125168.
43. Speiser, J. L., Miller, M. E., Tooze, J., Ip, E., 2019. A comparison of random forest variable selection methods for classification prediction modeling. Expert Syst. Appl. 134, 93-101.
44. Tang, Y., Horikoshi, M., Li, W., 2016. ggfortify: unified interface to visualize statistical results of popular R packages. R J. *8*(2), 474.
45. Thompson, J. A., Tan, J., Greene, C. S., 2016. Cross-platform normalization of microarray and RNA-seq data for machine learning applications. PeerJ. 4, e1621.
46. Vaughan, M. M., Christensen, S., Schmelz, E. A., Huffaker, A., Mcauslane, H. J., Alborn, H. T.,Teal, P. E., 2015. Accumulation of terpenoid phytoalexins in maize roots is associated with drought tolerance. Plant Cell Environ. *38*(11), 2195-2207.
47. Wang, H., Cimen, E., Singh, N., Buckler, E., 2020. Deep learning for plant genomics and crop improvement. Curr. Opin. Plant Biol. 54, 34-41.
48. Yu, G., 2020. Using ggtree to visualize data on tree-like structures. Curr. Protoc. Bioinform. 69(1), e96.
49. Yu, S. M., Lu, C. H., Ho, T. H. D., Shuen-Fang, L. O., 2018. U.S. Patent Application No. 15/838,702.

Tables

Table No. 1 List of genes identified through Boruta analysis

Sl. No.	Probe ID	Locus Id	Function
1	Os.15784.1.A1_s_at	LOC_Os05g43860.1	Expressed Protein
2	Os.27688.1.A1_at	LOC_Os04g21350.1	Flowering promoting factor-like 1
3	Os.39552.1.A1_s_at	LOC_Os06g48300.1	Protein phosphatase 2C
4	Os.54874.1.S1_at	LOC_Os09g27040.1	GEX1
5	Os.9653.1.S1_at	LOC_Os05g04360.1	Protein phosphatase 2C
6	OsAffx.12370.2.S1_at	LOC_Os02g36200.1	Expressed Protein
7	OsAffx.3920.1.S1_s_at	LOC_Os04g27670.1	Terpene synthase family
8	OsAffx.6874.1.S1_at	LOC_Os10g35340.1	Expressed Protein

Figures

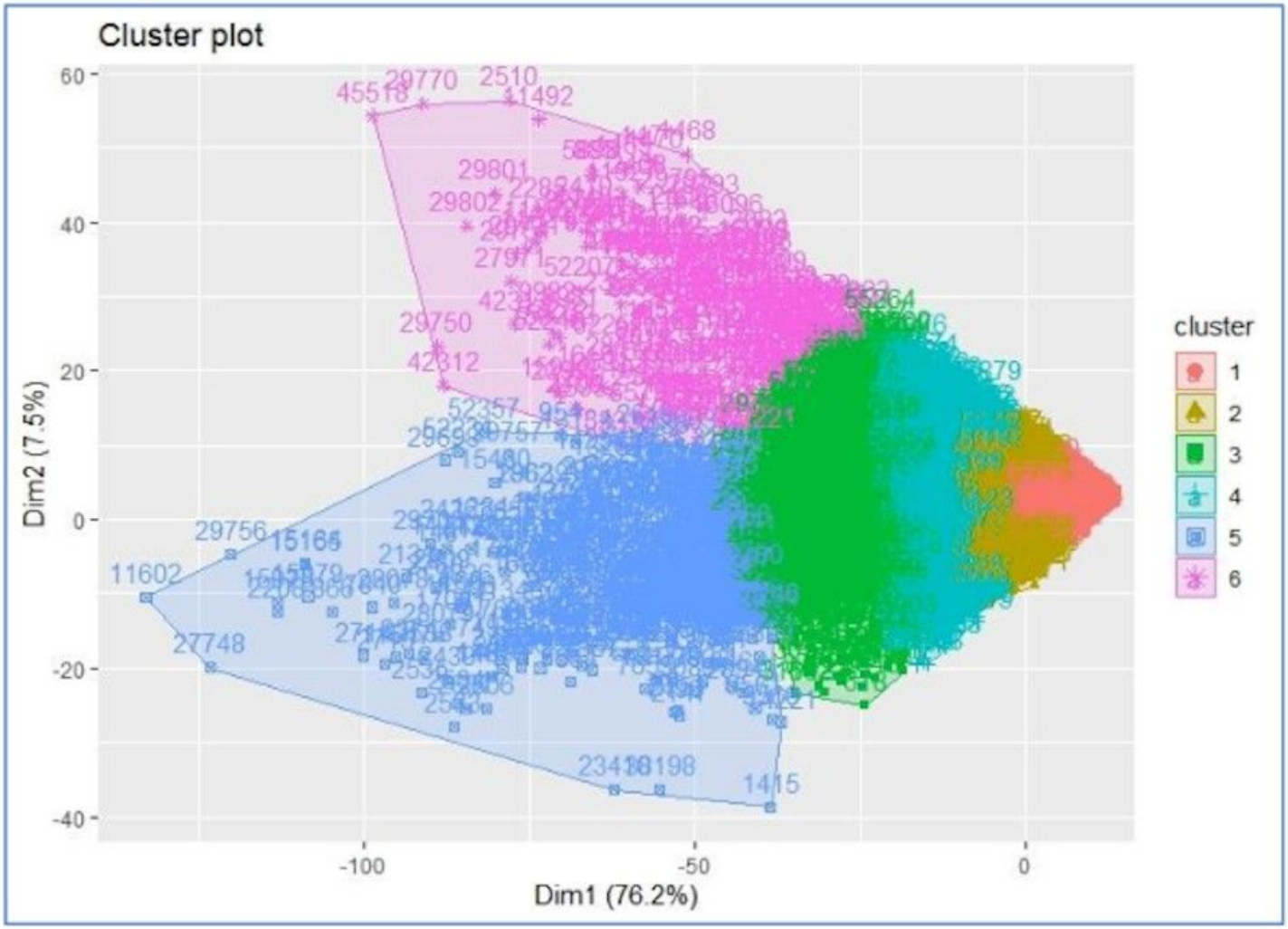

Figure 1

Clusters of genes identified using K-means clustering analysis

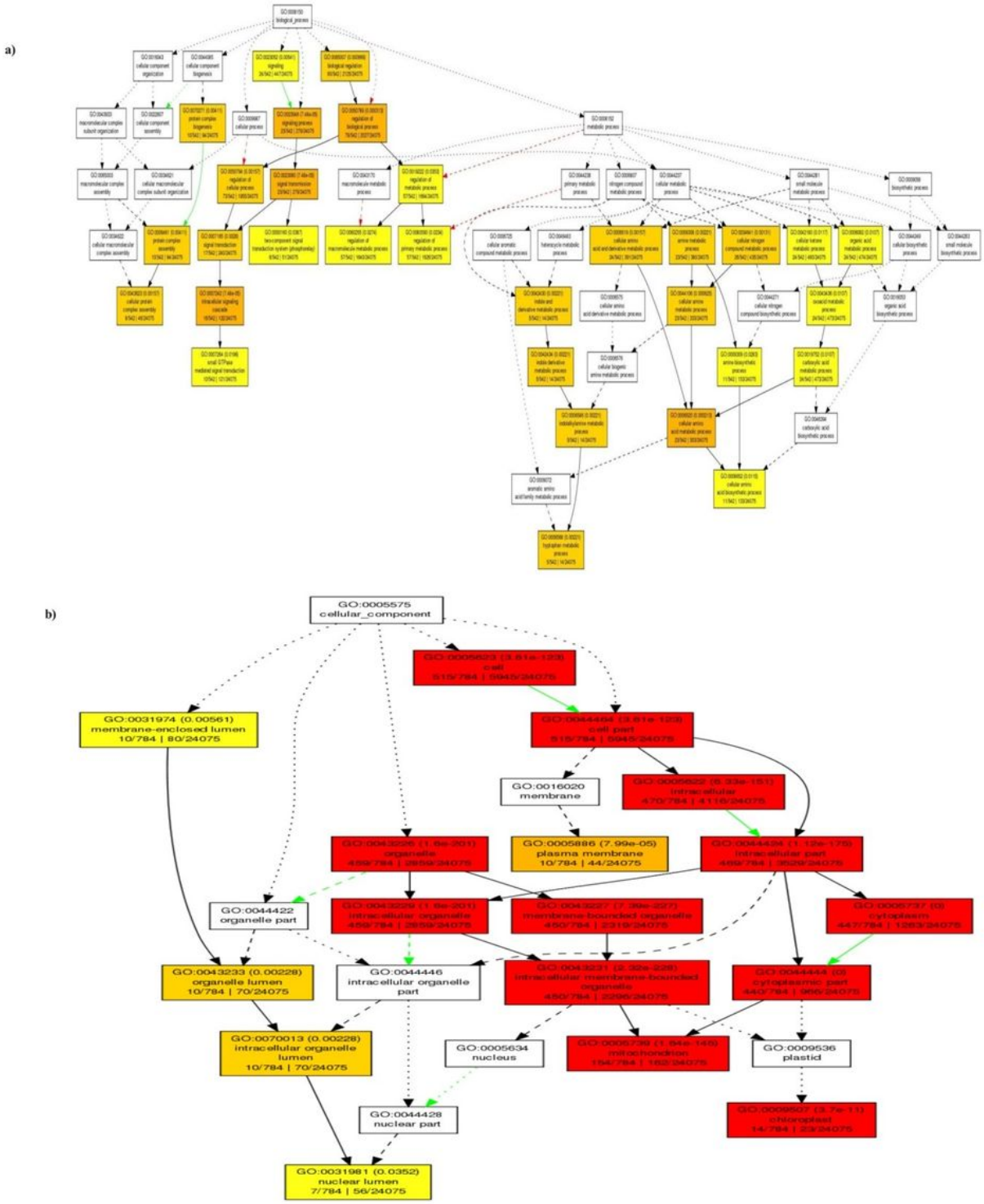

Figure 2

a. Gene ontology analysis of selected genes (Biological and metabolic processes b. Gene ontology analysis (Cellular processes)

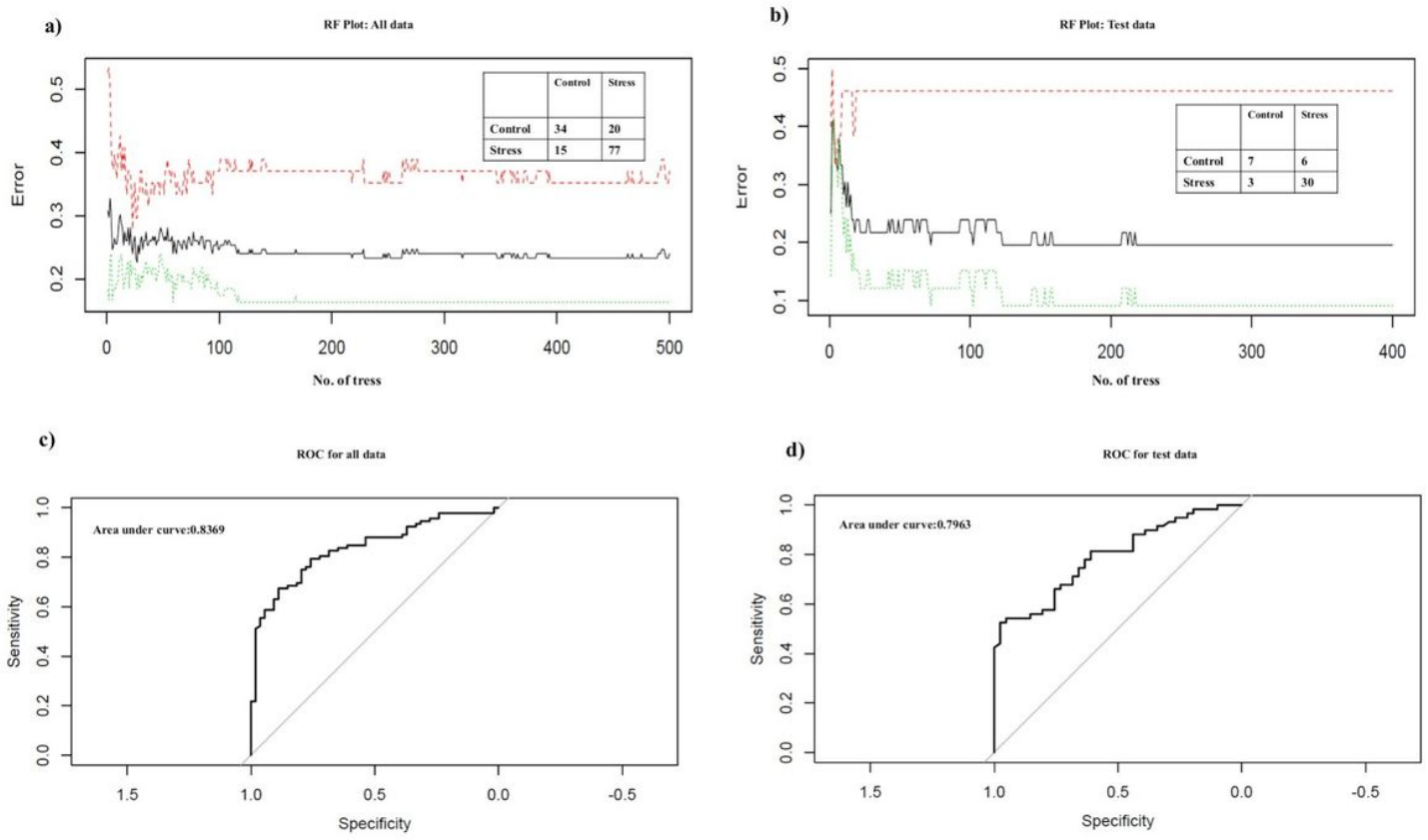

Figure 3

Random forest plot and ROC curve. a) Random forest plot for all the samples, b) Random forest plot for test samples, c) ROC curve for all samples, d) ROC curve for test samples. Tables in Fig. 3a and 3b is the class classification matrix of all samples and test samples, respectively.

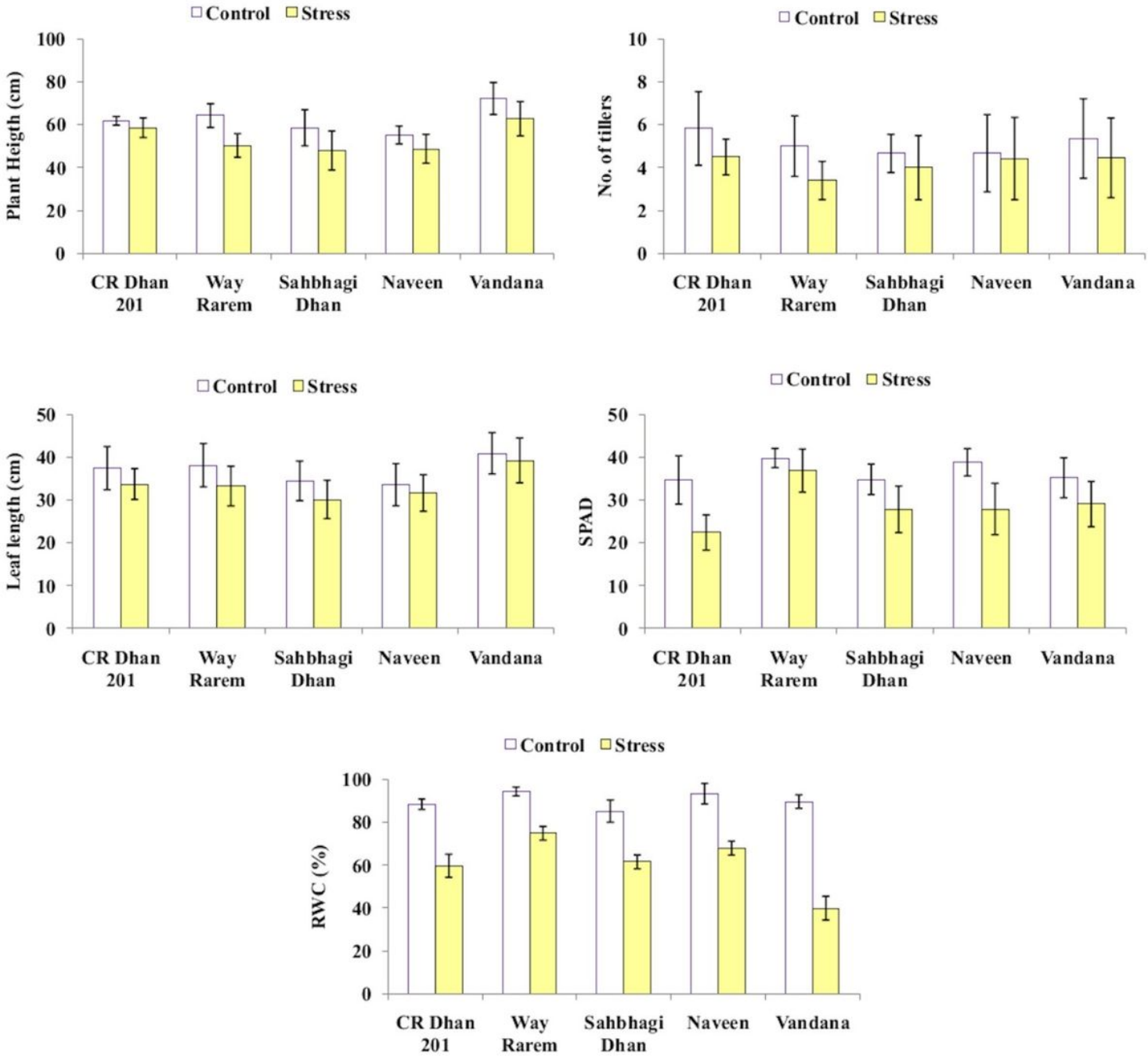

Figure 4

Morphological and physiological response of different rice varieties under drought stress.

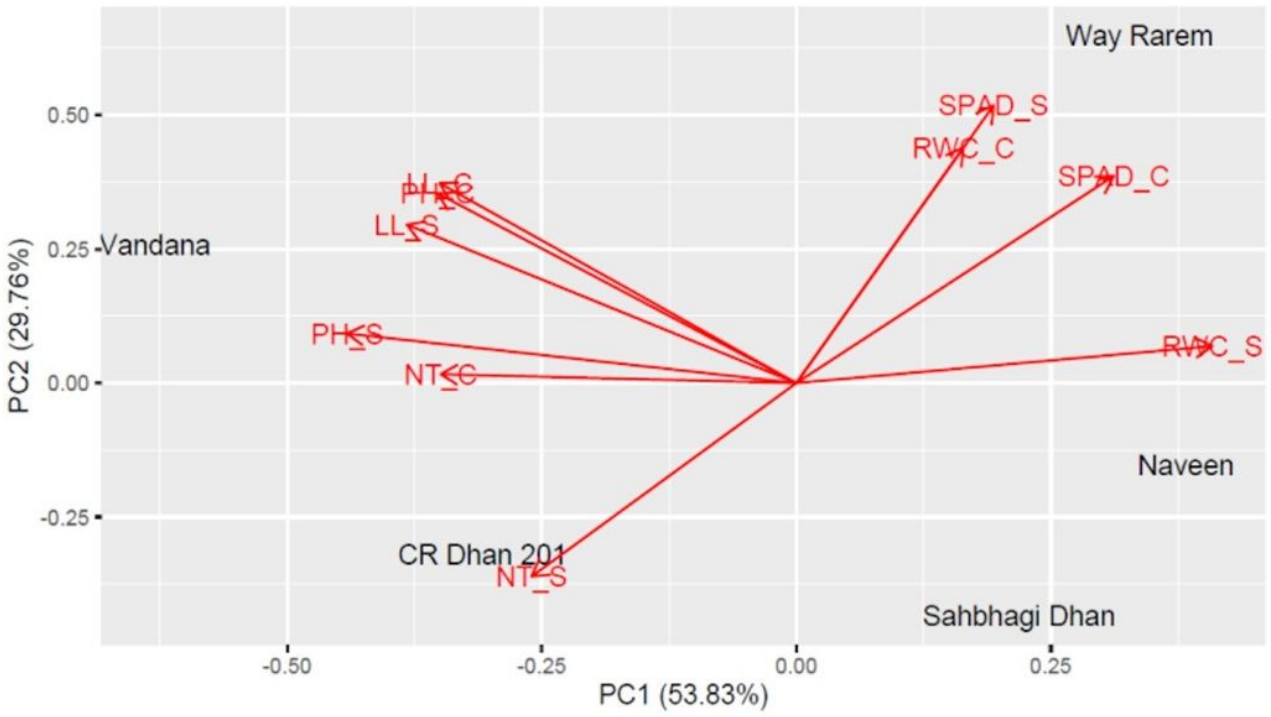

Figure 5

Principal component analysis of varieties and traits under control and stress condition. Variety_C and Variety_S indicates control and stress. RWC-Relative Water Content; SPAD-Cholorphyll index, LL-Leaf Length, NT-No. of tillers; PH-Plant Height

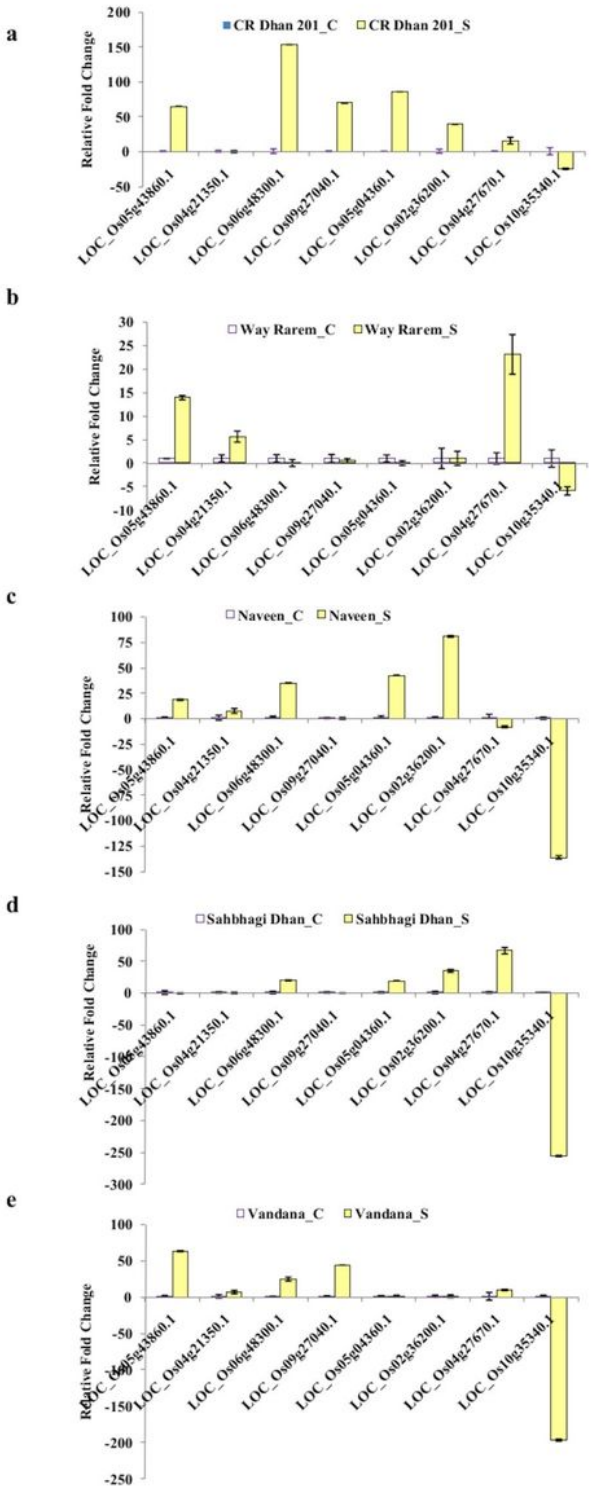

Figure 6

Expression analysis of eight genes in different varieties under drought stress condition. a) CR Dhan 201, b) Way Rarem c) Naveen d) Sahbhagi Dhan e) Vandana

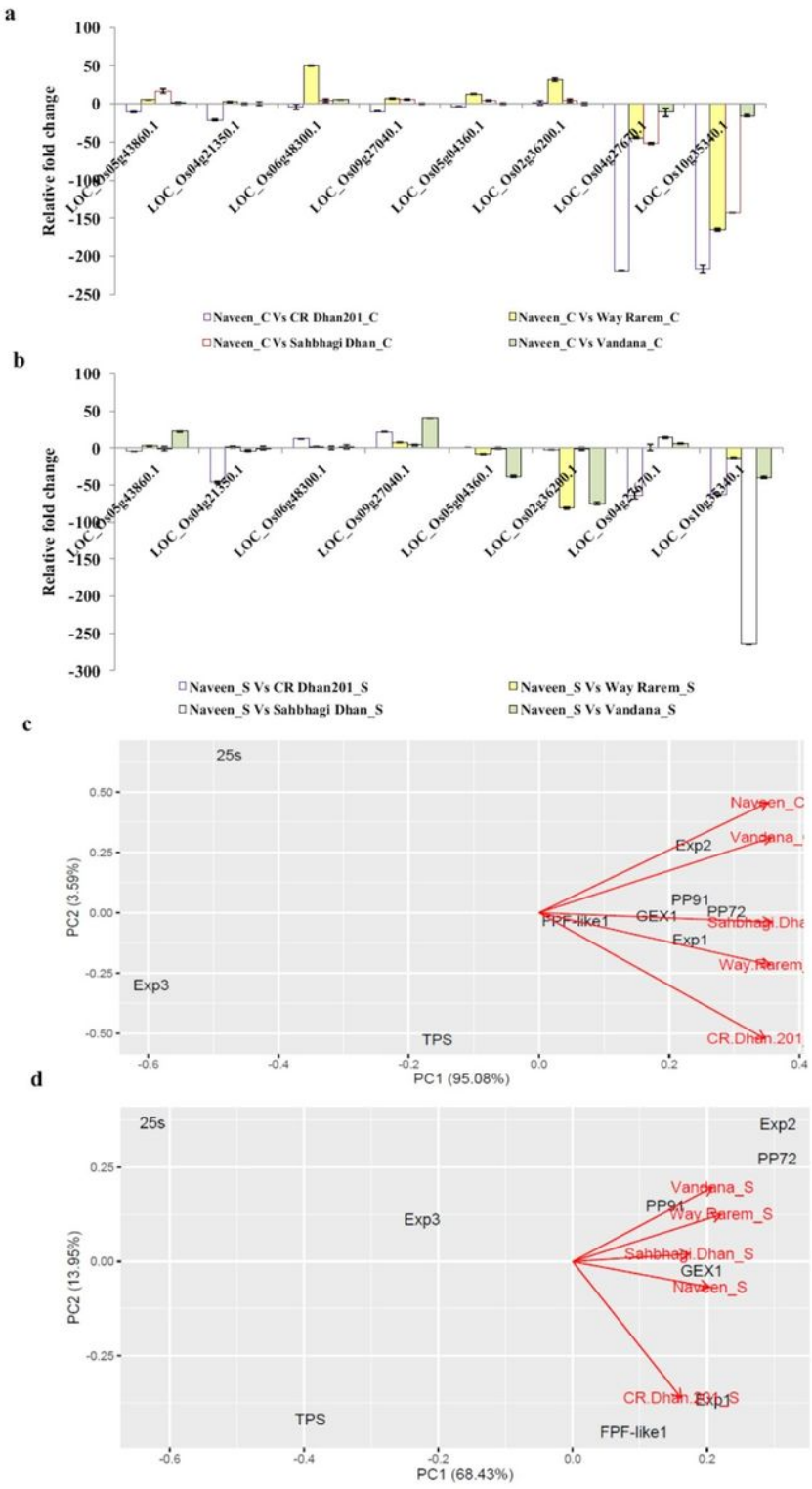

Figure 7

a PCA analysis of gene expression and varieties under control condition. b. PCA analysis of gene expression and varieties under stress condition

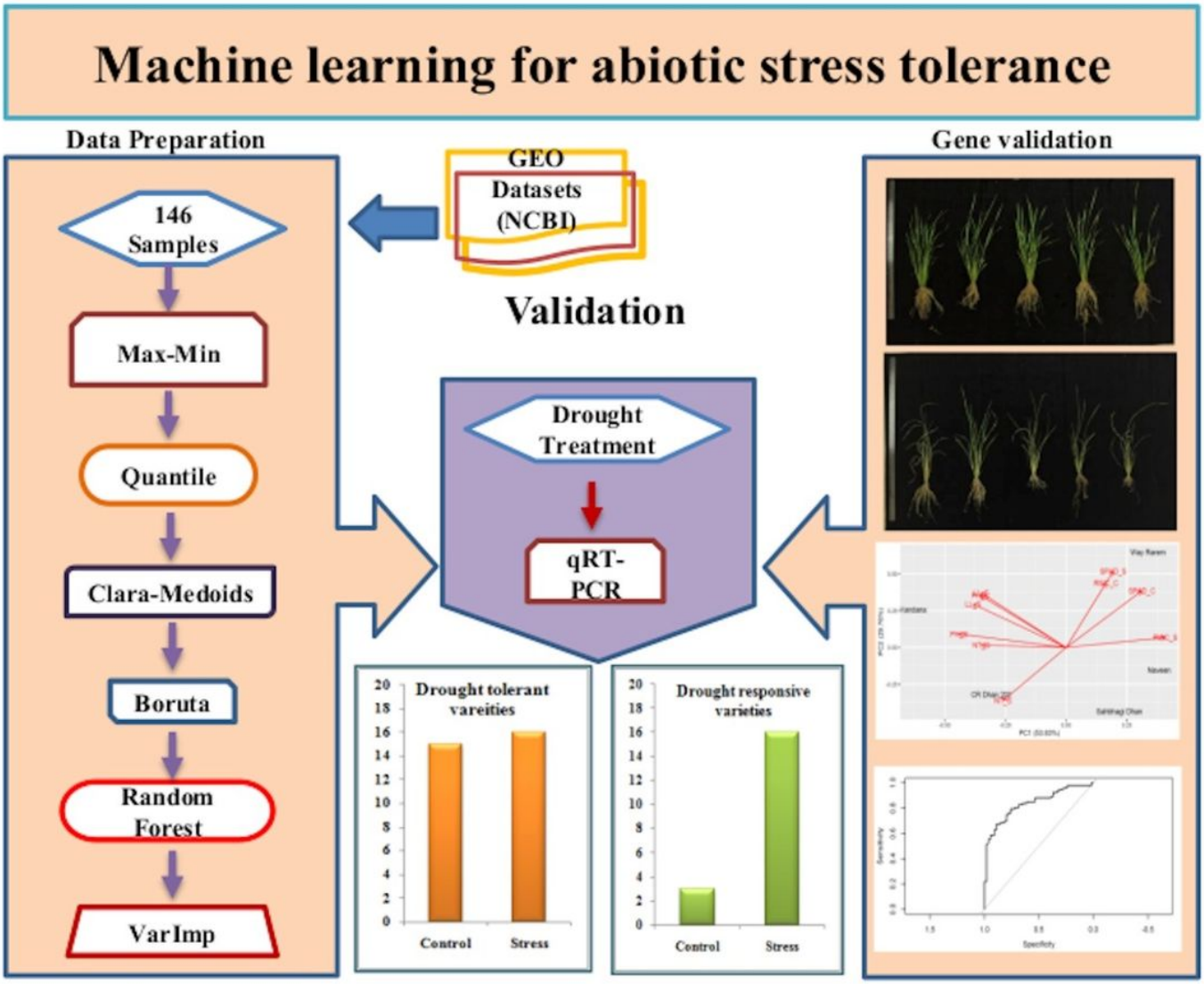

Figure 8

Machine learning approachused for the identification of stress responsive genes in rice

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTablesandfigures.docx