Running Head: Machine classification of multiple stresses in rice

Corresponding author: Wusirika Ramakrishna

Mailing address: Department of Biological Sciences, Michigan Technological University, 1400 Townsend Drive, Houghton, MI 49931, USA
Tel.: +1 906 487 3068; fax: +1 906 487 3167; E-mail: wusirika@mtu.edu
Machine Learning Approaches Distinguish Multiple Stress Conditions using StressResponsive Genes and Identify Candidate Genes for Broad Resistance in Rice

Rafi Shaik and Wusirika Ramakrishna

Department of Biological Sciences, Michigan Technological University, 1400 Townsend Drive, Houghton, MI 49931, USA

One sentence summary

Meta-analysis of differentially expressed rice genes under different stress conditions accurately classified them using machine learning approaches and identified genes likely to confer broad resistance to multiple abiotic and biotic stresses
Abstract

Abiotic and biotic stress responses are traditionally thought to be regulated by discrete signaling mechanisms. Recent experimental evidence revealed a more complex picture where these mechanisms are highly entangled and can have synergistic and antagonistic effects on each other. In the present study, we identified shared stress responsive genes between abiotic and biotic stresses in rice by performing meta-analyses of microarray studies. About 70% of the 1377 common Differentially Expressed Genes (DEGs) showed conserved expression status and majority of the rest were downregulated in abiotic stresses and upregulated in biotic stresses. Using dimension reduction techniques, Principal Component Analysis (PCA) and Partial Least Squares Discriminant Analysis (PLS-DA), we were able to segregate abiotic and biotic stresses into separate entities. The supervised machine learning model, Recursive-Support Vector Machine (R-SVM) could classify abiotic and biotic stresses with 100% accuracy using a subset of DEGs. Further, using Random Forests (RF) decision tree model, 8 out of 10 stress conditions were classified with high accuracy. Comparison of genes contributing most to the accurate classification by PLS-DA, R-SVM and RF revealed 196 common genes with a dynamic range of expression levels in multiple stresses. Functional enrichment and co-expression network analysis revealed the different roles of transcription factors and genes responding to phytohormones or modulating hormone levels in regulation of stress responses. We envisage the top ranked genes identified in this study which highly discriminate abiotic and biotic stresses as key components to further our understanding of the inherently complex nature of multiple stress response in plants.
The need to breed robust and high productivity crops is more important than ever due to increasingly adverse environmental conditions and scarce natural resources. Food productivity has to be raised by as much as 70-100% to meet the nutritional needs of the growing population which is expected to rise to 9 billion by 2050 (Godfray et al., 2010; Lutz and Samir, 2010). Rice is both a major food crop accounting for 20% of daily calorie intake of about 3.5 billion people (IRRI), and a model organism which shares extensive synteny and collinearity with other grasses. Thus, development of rice that can sustain a wide variety of adverse conditions is vital to meet the imminent global energy demands.

A broad range of stress factors divided into two major categories namely abiotic stresses encompassing a variety of unfavorable environmental conditions such as drought, submergence, salinity, heavy metal contamination or nutrient deficiency and, biotic stresses caused by infectious living organisms such as bacteria, virus, fungi or nematodes negatively affect productivity and survival of plants. Advancements in whole genome transcriptome analysis techniques like microarrays and RNA-seq have revolutionized the identification of changes in gene expression in plants under stress, making it possible now to chart out individual stress specific biomolecular networks and signaling pathways. However, in field conditions, plants are often subjected to multiple stresses simultaneously, requiring efficient molecular mechanisms to perceive multitude of signals and to elicit a tailored response (Sharma et al., 2013). Increasing evidence from experimental studies suggests that the cross talk between individual stress-response signaling pathways via key regulatory molecules, resulting in the dynamic modulation of downstream effectors’ is at the heart of multiple stress tolerance. A number of studies have identified many genes especially transcription factors and hormone response factors that play a central role in multiple stresses and manifest a signature expression, specific to the stress condition. For example, ABA response factors are upregulated in majority of abiotic stresses activating an oxidative response to protect cells from ROS damage but were found to be downregulated in a number of biotic conditions possibly suppressed by immune response molecules (Cao et al., 2011).

The wide range of abiotic and biotic stress factors and their numerous combinations in natural conditions generate a customized stress response. This suggests identification and characterization of key genes and their co-expression partners which show an expression profile that discriminates abiotic and biotic stress responses would increase our understanding of plant
stress response manifold and provide targets for genetic manipulation to improve their stress tolerance. The availability of multiple genome-wide transcriptome data sets for same stress condition provides an opportunity to identify, compare and contrast stress specific gene expression profile of one stress condition with other stresses. Meta-analysis by combining similar studies provides a robust statistical framework to reevaluate the original findings, improve sensitivity with increased sample size and to test new hypotheses. Meta-analysis of microarray studies is widely used especially in clinical research to improve statistical robustness and detect weak signals (Liu et al., 2013; Rung and Brazma, 2013). For instance, thousands of samples belonging to hundreds of cancer types were combined which provided new insights into the general and specific transcriptional patterns of tumors (Lukk et al., 2010). Microarray studies are burdened with high dimensionality of feature space also called as ‘curse of dimensionality’ i.e. availability of very many variables (genes) for very few observations (samples). Machine learning algorithms (supervised and unsupervised) such as Principal Component Analysis (PCA), decision trees and Support Vector Machines (SVM) provide a way to efficiently classify two or more classes of data. Further feature selection procedures like Recursive-SVM (R-SVM) provide means to identify the top features contributing most to the accuracy of classification.

In the present study, we performed meta-analysis of stress response studies in rice using publically available microarray gene expression data conducted on a single platform (AffymetrixRiceArray). Meta-analysis of abiotic and biotic stresses was performed separately to identify differentially expressed genes involved in multiple stress conditions. The lists of abiotic and biotic DEGs were then compared to identify common genes with conserved and non-conserved gene expression i.e. whether up or down or oppositely regulated in both the categories, revealing the broad patterns of their involvement in stress response. In order to test the efficiency of identified common DEGs in classification of abiotic and biotic stresses as well as individual stresses within abiotic and biotic stresses, we systematically investigated various classification and machine learning techniques including PCA, Partial Least Squares Discriminant Analysis (PLS-DA), SVM and Random Forest (RF). We characterized the shared DEGs through functional enrichment analysis of gene ontologies, metabolic pathways, transcription factor families and microRNAs targeting them. We also analyzed correlation of co-expression between the common DEGs to find sets of genes showing high co-expression and identify hub genes which show most number of edges over a very high cut-off value.
RESULTS

Differentially Expressed Genes Common to Abiotic and Biotic Stresses

We analyzed 559 microarray samples (219 from abiotic and 340 from biotic stresses) from 13 stress conditions of which 7 were abiotic (cold, drought, heat shock, metal, nutrient, salt and submergence) and 6 were biotic stresses (bacteria, fungi, insect, nematode, virus and weed) (Supplemental Table S1A). Meta-analysis by combinatorial analysis of 7 abiotic stresses from 15 different studies together identified 3471 DEGs and 6 biotic stresses from 17 different studies revealed 3065 DEGs with false discovery rate (FDR) \( \leq 0.01 \) (Fig. 1A and Supplemental Table S2). About 60% of DEGs in abiotic stresses were downregulated while 60% of DEGs in biotic stresses were upregulated (Fig. 1B). This broad pattern indicates that a wide variety of biological processes are downregulated under abiotic stress as it affects the whole system thus driving the plant to a protective and energy conserving mode. On the other hand, biotic stresses are often localized especially at the early stages and require an array of defense response molecules and metabolites to be synthesized and orchestrated as in for example systemic acquired resistance (SAR) to execute a resistance response against a specific infectious organism. Among the DEGs, more than 26% or 1377 genes were common to abiotic and biotic stresses indicating that these genes which are just 3.5% of all non-TE genes in rice (MSU7.0) are affected by a diverse set of stress conditions and possibly play significant roles in multiple stress responses (Supplemental Table S3). Our major objective in this study is to analyze the stress responsive genes involved in multiple stresses that regulate cross talk between abiotic and biotic stresses. Therefore, we focused on the 1377 common DEGs for our study.

We found 72% or 999 out of 1377 common DEGs with conserved expression between abiotic and biotic stresses suggesting most of these genes and their associated biological processes are regulated in a similar fashion in vast majority of stress conditions. Among the 28% of DEGs showing non-conserved expression, 21% or 295 genes were downregulated in abiotic and upregulated in biotic stress (Fig. 1C). About 16% or 221 of these genes are annotated as ‘expressed protein’ and ~7% or 96 have no GOSlim assignment revealing that many of stress
responsive genes are still poorly understood. Studies elucidating functional roles of these genes would be vital for comprehensive understanding of stress response in rice.

**Machine Learning Approaches based on Common DEGs Classified Abiotic and Biotic Stresses into Two Classes with High Accuracy**

We investigated if the different stress conditions can be accurately classified using the common DEGs employing machine learning approaches. Initially, we investigated the performance of PCA in discriminating abiotic stresses from biotic stresses as two-classes using all of the 1377 common DEGs. The first three PCs captured 56.4% of variance between the samples. The 3D-PCA plot of top 3 PCs showed clear separation of abiotic and biotic classes for majority of the samples although both the classes were widely dispersed across components (Fig. 2A). Nonetheless, there were some samples showing considerable overlap between the classes. We then analyzed the data-set using Partial Least Squares Discriminant Analysis (PLS-DA), a technique that is specifically suited for analysis of data-set with high feature dimensions and multicollinearity (Perez-Enciso and Tenenhaus, 2003). Many of the published microarray studies have found PLS-DA as a highly efficient method for multiclass classification (Student and Fujarewicz, 2012). PLS-DA resulted in five components which captured ~62% variance between the two classes and separated them with a very high accuracy of 0.99 (R²:0.95 (goodness of fit), Q²: 0.93 (predictive value) p-val<0.01) upon 10 fold cross-validation. The 3D plot of PLS-DA showed clear separation of all the samples between abiotic and biotic stresses (Fig. 2B). The important genes contributing most to the PLS-DA separation can be identified using Variable Importance in Projection (VIP) score which is a weighted sum of squares of PLS loadings (Perez-Enciso and Tenenhaus, 2003). There were 177 genes with the VIP-score (component 1) cutoff value ≥1.5 (Zhang et al., 2013) and 33 genes with values ≥2 (Supplemental Table S4).

Next, we analyzed the same data-set using another very popular supervised learning technique for microarray data classification called Recursive-Support Vector Machine (R-SVM) which identified 540 genes (39.2% out of 1377) that can classify abiotic and biotic stresses with 100% accuracy and 88 (6%) genes with 95% accuracy after rigorous cross-validation using LOOCV (leave one out cross validation) (Fig. 3). These 540 genes included a number of hormone response and stress response signaling genes. All five of the MYB TFs which are important regulators of development and defense responses in plants (Yanhui et al., 2006) found
in the common DEGs were part of these 540 genes. Further, 103 (19%) of the 540 genes were part of a recently published database namely stress-responsive transcription factor database (STIFDB2) (Naika et al., 2013) which provides a list of stress responsive genes (1118 genes of *Oryza sativa* subsp. *japonica*) identified through biocuration and genomic data mining. Out of 540 genes, 178 (33%) were the ones with non-conserved expression pattern between abiotic and biotic stresses, which is slightly higher compared to the 28% of genes showing non-conserved expression in all of the common DEGs. Although PCA based on these 540 genes resulted in poor separation of the classes with 47.4% variance captured by top 3 PCs, PLS-DA showed clear separation of the two classes (Fig. 2C and D). The top 5 components of PLS-DA captured 53% of variance with classification accuracy of 0.97 ($R^2$:0.91, $Q^2$:0.87 p-val<0.01) which is slightly less than the 0.99 accuracy obtained using all 1377 common DEGs. There were 79 genes (14% of 540) with the VIP-scores $\geq$1.5 and 27 genes with $\geq$2. There were two genes with VIP-scores $\geq$3 which code for xylanase inhibitor and glycosyl hydrolase both showing conserved upregulation.

**Analysis of Shared DEGs Identified Top Genes with Discordant Behavior among Multiple Stresses**

From the 13 stress conditions analyzed, we selected top 10 stresses (5 abiotic stresses: drought, metal, salt, cold and nutrient and 5 biotic stresses: bacteria, fungus, insect, weed and nematode) based on higher number of microarray samples. We analyzed this data using the normalized and pareto scaled intensities of 1377 DEGs to assess the performance of these genes for classification of different stress conditions. The top five components of PLS-DA captured 62.9% of variance between various stresses and showed classification accuracy of 0.77 ($R^2$:0.92, $Q^2$: 0.88 p-val<0.01). There were 196 and 53 genes with VIP scores (component 1) $\geq$1.5 and $\geq$2. The relatively low classification accuracy reflects the inherent similar expression patterns between different stresses. Nonetheless, the components 1 and 3 as shown in the 2D score plot and top three components as shown in 3D score plot were able to clearly separate abiotic and biotic stresses as two major groups (Fig. 4). The 2D and 3D plots also showed wide dispersion of drought stress and closeness with majority of cold stress samples. Similarly, the 3D plot showed higher overlap between salt and metal stresses than other stresses suggesting higher similarity of gene expression profile between them. The nutrient stress samples can be observed as a distinct
group although closer to other abiotic stresses. Bacterial stress samples show two major groups. One of the groups with most of bacterial samples showed overlap with fungal stress samples only. The other group was closer to weed, nematode and fungal stress samples. Insect stress was observed as a distinct group closer to the group with bacterial and fungal samples.

The same data-set was analyzed using another classification technique called Random Forest (RF) which classified 8 of the 10 stresses with 100% accuracy with an overall out-of-box (OOB) error rate of 0.0087 which is an unbiased estimate of classification error based on the one third left out samples (test samples) after bootstrap sample selection (Table I). Two of the stresses with less than 100% accuracy of classification were salt with one wrongly classified sample (error rate: 0.037) and fungal stress with two wrongly classified samples (error rate: 0.08). RF also provides a measure of variable importance by evaluating the increase in OOB error rate upon permutations called mean decrease in accuracy (Hsueh et al., 2013). The top 15 significant genes based on mean decrease in accuracy are shown in figure S2 including, LOC_Os02g45170 (error rate: 0.0056), a bHLH TF and LOC_Os05g31040, which codes for cytokinin dehydrogenase precursor.

**Functional Enrichment Analysis Revealed Molecular Mechanisms and Gene Families in Conserved and Non-conserved Gene Sets**

Gene ontology enrichment analysis of the 560 genes showing conserved downregulation in abiotic and biotic stresses identified major biological and cellular processes including photosynthesis (FDR: 1.40E-07), electron carrier activity (FDR: 3.60E-06), small molecule biosynthetic process (FDR: 2.10E-05), cellular nitrogen compound metabolic process which is the parent term for a number of amino acids and nucleobase-containing compounds to be overrepresented. The terms transcription repressor activity (FDR: 0.0008) and response to oxidative stress (FDR: 0.034) were also found to be significant (Fig. S3 and Supplemental Table S4). On the other hand, 439 genes showing conserved upregulation revealed a number of terms related to regulatory processes. The most significant innermost child terms are serine-type endopeptidase inhibitor activity (FDR: 2.2E-06) and chitin catabolic process (FDR: 0.00013). Serine proteases perform diverse physiological roles in plants, important among which are induction after pathogen attack leading to hypersensitivity response (HR), regulation of rubisco proteolysis, stomata development, perception of growth hormones, symbiosis and senescence
(Antao and Malcata, 2005; van der Hoorn, 2008). Significant enrichment of inhibitors of serine-type endopeptidases in diverse stress conditions indicates induction of several activities repressed by serine proteases as part of stress response. Further, serine protease inhibitors were also found to act as defense proteins by suppressing the activity of bowel proteinases in insects and plant pathogenic microorganisms (Mosolov and Valueva, 2011). Among the genes showing non-conserved expression, the set of genes downregulated in abiotic stresses and upregulated in biotic stresses were enriched with GO terms, which include extracellular region (FDR: 5.30E-06) and catalytic activity (5.3E-05).

Metabolic pathway enrichment analysis by DAVID revealed a number of pathways specifically enriched in one of the sets. In the conserved downregulated gene set, there were four annotation clusters with enrichment score >2.0 related to porphyrin and chlorophyll metabolism, transcription repressor activity via Nmr-A like domain which is involved in post-translational modification of the GATA-transcription factors (Stammers et al., 2001), photosynthesis and nicotianamine synthase activity. There were three annotation clusters with enrichment score >2.0 in conserved upregulated gene set related to heat shock protein Hsp20, valine, leucine and isoleucine degradation, and Bowman-Birk proteinase inhibitor (BBPI) family of serine protease inhibitors. In rice, BBPI genes were reported previously to be induced in multiple stresses like wounding, infection and hormonal stress (Rakwal et al., 2001; Qu et al., 2003). The top annotation clusters in the non-conserved abiotic down and biotic up gene set were made up of a number of interpro domain terms, glycoprotein, metal-ion binding, plant peroxidases and glycoside hydrolases.

There were 97 transcription factor and regulator genes in the common DEGs (7%) belonging to 24 gene families which showed a distinct pattern (Supplemental Table S5). The major TF families NAC, HSF, WRKY, MYB, and MYB_related were part of conserved downregulated genes and non-conserved genes downregulated under abiotic and upregulated under biotic stresses. Similarly, the TF families, ERF, bZIP, bHLH and three others were part of conserved upregulated genes and/ or non-conserved genes upregulated under abiotic and downregulated under biotic stresses. Twelve out of thirteen Ethylene Response Factors (ERFs) were found in conserved upregulated gene sets. These AP2 (APETALA2) domain containing ERFs are well known for their role in both abiotic and biotic stress responses and were also shown to enhance multiple stress tolerance (Xu et al., 2011). Nine out of twelve WRKY TFs
were part of non-conserved genes downregulated under abiotic and upregulated under biotic stresses which suggests that these TFs are the major regulatory factors that determine the direction of molecular machinery and ultimately the cellular fate under simultaneous multiple stresses. MYB along with NAC TFs are reported to control antagonism between hormone-mediated abiotic stress and pathogen response pathways (Atkinson and Urwin, 2012). On the other hand, all five of G2 (Golden2)-like TF family members which also contain MYB-like DNA binding domain were part of conserved upregulated gene set. The G2-like TFs are required for chloroplast development and were shown to influence nuclear photosynthetic gene expression (Waters et al., 2009). We found a dearth of studies on the role of G2-like TFs under stress conditions. Downregulation of photosynthetic mechanisms under stress is well established as also observed in the enriched GO terms in our conserved downregulated gene set. Genetic manipulation of G2-like TFs would shed further light on regulation of photosynthesis under stress and reveal novel mechanisms to enhance stress tolerance. Out of the five LSD (Lesion Simulating Disease) (Dietrich et al., 1997) family members reported in *Oryza sativa* subsp. *japonica* by PlnTFDB, two were part of conserved downregulated gene set. LSD TFs act as negative regulators of programmed cell death (PCD) in a hypersensitive response (HR) (Epple et al., 2003). Transgenic suppression of LSD orthologs in rice resulted in a dwarf phenotype due to deficiency of bioactive gibberellin while overexpression of LSD enhanced resistance to rice bacterial blight (Xu and He, 2007). Based on our finding, studying LSD TFs under simultaneous abiotic and biotic stresses would provide vital clues on stress cross-talk and modulation of PCD.

We analyzed the microRNAs predicted to target the 1377 common DEGs using the database PMRD (Zhang et al., 2010). Out of the 456 experimentally verified miRNAs (miRBase (Griffiths-Jones et al., 2006)) in rice, 142 (31%) miRNAs belonging to 50 miRNA families were found to target one or more common DEGs (Supplemental Table S6). Recently, 35 miRNAs from 31 miRNA families were found to be differentially expressed under abiotic stresses, drought, salt and cold (Shen et al., 2010). Eighteen of these 31 stress responsive miRNA families were part of the 50 miRNA families targeting the common DEGs. The miRNA osa-miR1436 was found to target five of the conserved upregulated genes including LOC_Os09g23620, a MYB TF while osa-miR446 was found to target five of the conserved downregulated genes.
Co-expression Analysis Revealed Two Dense Clusters of Positively and Negatively Correlated Genes under Multiple Stresses

We conducted co-expression analysis using the normalized gene expression values of the common DEGs from stressed microarray samples and calculating Pearson Correlation Coefficient \( r \) between them. Out of the 947,376 possible edges (co-expression gene pairs) between the common DEGs, we found 8,924 edges with very high correlation \( r \geq 0.9 = 4,254 \) and \( r \leq -0.7 = 4,670 \) edges, \( p\text{-value} = 0.01 \) in abiotic stress samples and 21,229 edges \( r \geq 0.9 = 7,673 \) and \( r \leq -0.7 = 13,656 \) edges, \( p\text{-value} = 0.01 \) in biotic stress samples. A very high number of negative edges were observed in biotic stresses compared to abiotic stresses. For instance, there were 88 edges in biotic stresses with \( r \leq -0.9 \) but only four edges in abiotic stresses with \( r \leq -0.9 \). There were 3,701 shared edges between the two data-sets with \( r \geq 0.9 \) and \( r \leq -0.7 \), out of which 2,684 (72\%) were positive edges and 1,017 were negative edges. These 3,701 edges were between 381 genes, out of which 257 (67\%) genes showed conserved downregulation, 54 genes showed conserved upregulation and 49 genes showed downregulation in abiotic stresses and upregulation in biotic stresses. The 2,684 positive edges were between 208 genes, out of which 194 (93\%) genes showed conserved downregulation. Among the 381 genes, 15 had >75 high correlation edges. The top three genes with most number of edges were, LOC_Os02g22480 (glycosyltransferase -142 edges), LOC_Os11g47840 (putative rhomboid homologue - 120) and LOC_Os03g57200 (glutathione S-transferase - 93). All three of these genes showed conserved upregulation. Among the 14 TFs with significant edges, three TFs belonging to NF-Y (Nuclear Factor –Y, a histone like CCAAT-binding domain TF), G2-like and bHLH TF families had most number of significant edges (79, 37 and 20, respectively). Majority of these edges were positive edges with other genes that showed conserved downregulation.

We analyzed the 3,701 significant edges using the plugin NetworkAnalyzer in network analysis platform Cytocape 2.8.3 (Shannon et al., 2003) which revealed a dense cluster of positive edges (edges with \( r \geq 0.95 \) are shown in red color) which included most of the nodes with >75 edges (shown in blue) and a sparse cluster of negative edges (edges with \( r \leq -0.9 \) are shown in green) (Fig. 5). The two positive edge and negative edge rich clusters were found to be bridged by the gene LOC_Os01g13570, coding for phosphoglyceratemutase with a positive edge to SOUL heme-binding protein that was highly connected to negative edge rich cluster and
positive edges with rhodanese and pentatricopeptide (PPR) domain containing proteins which were highly connected to the positive edge rich cluster.

**High Overlap among Genes Identified by Different Classification Techniques, Co-expression and Functional Enrichment Analysis**

We compiled the significance of the common DEGs based on various criteria including feature importance identified by different classification techniques, count of number of co-expression edges, PlnTFDB gene, and STIFDB2 gene (Supplemental Table S3). We found that many of the PLS-DA two-class significant genes (177 genes with VIP ≥ 1.5) were also significant in PLS-DA multiclass (36% or 71 out 196) and RF’s top 100 genes (68%) but showed poor overlap with the 540 significant genes found by SVM (2% or 9 out of 540), TF genes (9% or 9 out of 97) and STIFDB2 genes (10% or 27 out of 259). However SVM’s 540 genes showed high overlap with PLS-DA multiclass (50% or 99 out of 196), TF genes (45% or 44 out of 97) and STIFDB2 genes (40% or 103 out of 259). Taken together, the 196 top genes of PLS-DA multiclass showed overlap with most of the other significant gene lists, of which 43 (22%) were also part of STIFDB2 list. Out of 1,118 *Oryza sativa* subsp. *japonica* genes reported as stress responsive genes in STIFDB2 (Naika et al., 2013), 259 (23%) were part of common DEGs. Further, out of 97 TF genes in the common DEGs, only 12 were part of STIFDB2 and none of the major WRKY and MYB TF genes including those previously reported as stress responsive genes (Atkinson and Urwin, 2012) were part of STIFDB2’s list. The top 10 of these 196 genes are given in Table II. The topmost gene encodes a CCCH zinc finger domain containing TF known to control embryogenesis (Li and Thomas, 1998) and involved in multiple abiotic stresses (Sun et al., 2007; Kim et al., 2008). A homolog of this gene (LOC_Os05g10670) which was also part of the 1432 upregulated genes in our meta-analysis of abiotic stresses, was recently reported to confer delayed senescence and improved tolerance to high-salt and drought stresses by regulating reactive oxygen species homeostasis, and metal homeostasis (Jan et al., 2013). One gene which was part of all feature selection lists was LOC_Os11g26780, a dehydrin gene which had one significant positive edge with another dehydrin gene (LOC_Os11g26790, r=0.97 and 0.93 in abiotic and biotic stresses, respectively) both of which showed conserved upregulation.

Comparison of the common DEGs with the list of 1922 hormone related genes of *Arabidopsis* as reported in Arabidopsis Hormone Database 2.0 (Jiang et al., 2011) using putative
orthologous genes found by GreenPhylDB (Rouard et al., 2011) revealed 31 common DEGs that were orthologous to 51 *Arabidopsis* hormone genes (Supplemental Table S3). A summary table of the expression status of hormone related genes in the common DEGs (78 genes) based on Arabidopsis hormone database orthologs and paralogs with same annotation and expression status in both abiotic and biotic stresses (except TFs) or name of the hormone in the gene annotation provided by MSU 7.1 is given in Table III. Overall, the expression status of various hormone related genes was very similar to the one proposed in a recent review (Atkinson and Urwin, 2012). For instance, 9 out 12 abscisic acid responsive genes showed conserved upregulation while 6 out of 10 ethylene responsive genes showed non-conserved downregulation under abiotic and upregulation under biotic stress. Most of the conserved auxin downregulated genes were related to auxin biosynthesis and response factors while conserved upregulated were related to auxin repressed factors which indicates extensive downregulation of auxin induced biological processes. A recent study analyzed transcriptome of rice under bacterial stress by *Xanthomonas oryzae* pv. *oryzae* and compared the DEGs with those found in seven other microarray studies conducted on AffymetrixRiceArray (Narsai et al., 2013). They reported 240 genes (212 loci) as differentially expressed in multiple stresses. Out of these loci, 110 (51.8%) were part of our common DEGs list, most of which belonged to conserved upregulation gene set (64%) and included many important genes such as WRKY, AP2/EREBP TFs, ABC transporter, multidrug resistance and universal stress genes.

**DISCUSSION**

Multiple stress response in plants has been a hot topic of research as many studies, including those involving genetic manipulation and chemical intervention reported increased resistance to one stress resulted in heightened susceptibility to other abiotic and biotic stress conditions (Atkinson and Urwin, 2012; Sharma et al., 2013). Further, it was suggested that plant hormones are the key determinants of genetic switches and cellular adjustments in a multi-stress environment. Different plant hormones are broadly categorized to play central roles in different kinds of stress responses. For instance, within biotic stresses, (hemi)biotrophic pathogens commonly activate salicylic acid (SA)-dependent defense response, while necrotrophic pathogens activate jasmonic acid (JA) and ethylene (ET)-dependent signaling pathways (Sharma et al., 2013). SA and JA/ET often act antagonistically and propagate opposing influences
(Pieterse et al., 2009). On the other hand, abscisic acid (ABA) is well established as the major player of abiotic stress response. ABA is increasingly found to also play a critical role in biotic stresses by negatively regulating plant immunity. Many studies found that abiotic stresses enhance plant susceptibility to pathogen attacks due to weakening of defense systems. Thus, it was proposed that plants prioritize abiotic stress tolerance over biotic stress response with ABA as molecular switch between the two responses to minimize the damage (Lee and Luan, 2012). Recently, however, contrary studies where biotic stress takes precedence have been reported (Kim et al., 2011; Mang et al., 2012; Sanchez-Vallet et al., 2012). Thus, in light of these recent developments which revealed a rather complicated picture of multiple stress response, we embarked on identification of differentially expressed genes in abiotic and biotic stress environments separately and performed comparative analysis of the shared stress responsive genes, which would provide vital clues on the causative factors behind the cross-talk resulting in the observed synergistic and antagonistic regulation of known abiotic and biotic stress response pathways.

Our study identified 1377 differentially expressed common genes under a wide spectrum of abiotic and biotic stress conditions, their expression status can be considered as a representation of their overall involvement in stress response to non-living factors and living organisms. Thus, this list of genes forms an ideal gene set to objectively investigate the similarities and differences between abiotic and biotic stress responses. Although >70% of common DEGs showed conserved differential expression, we were able to classify different stresses including abiotic and biotic stresses with high accuracy indicating that their subtle expression differences can be exploited to effectively discriminate between various stress conditions.

A closer look at chloroplast and photosynthesis related genes in the common DEGs revealed conserved downregulation of 17 out of 18 photosystem II, chlorophyll A-B binding and thylakoid lumenal genes (Fig. 6). A diverse set of 40 chloroplast precursor enzymes which contain an amino-terminal transit peptide for import into chloroplast (Jarvis, 2008) were also part of the common DEGs, 26 (65%) of which showed conserved downregulation. Further, a number of cytochrome P450 genes (29 genes) which encode parent compounds for a number of secondary metabolites involved in plant defense (Jirschitzka et al., 2013) were part of the common DEGs. Fourteen of these 29 (~48%) genes showed conserved downregulation, while 7
showed conserved upregulation and the rest showed non-conserved differential expression. Thus, exploring the non-conserved DEGs would shed further light on the cross-talk of stress response via metabolic adjustments. Cell wall is the first line of plant defense in response to external stimuli. A number of important gene families involved in cell wall synthesis and modifications showed distinct patterns of expression under abiotic and biotic stresses. For instance, there were 6 OsWAK (Wall Associated Kinase) genes in common DEGs, all of which showed non-conserved downregulation under abiotic stresses and upregulation under biotic stresses (Fig. 6). WAKs are part of the transmembrane Receptor-Like Kinase (RLK) superfamily, which perceive stimuli using extracellular domains with signal transmission through their cytoplasmic kinase domains (Li et al., 2009). There are currently 144 rice genes regarded as WAKs (MSU7.0) compared to 26 genes in Arabidopsis which is most likely due to lineage specific gene duplications (Zhang et al., 2005). However, very little is known about the function of these genes in rice except OsWAK1 whose overexpression increased resistance to the blast fungus, Magnaporthe oryzae (Li et al., 2009; Kohorn and Kohorn, 2012). FAS1 (fasciclin-like) domain containing genes are another group of transmembrane genes involved in cell adhesion (Johnson et al., 2003; Ma and Zhao, 2010). All 5 of fasciclin domain genes in the common DEGs showed conserved downregulation. Similarly, most of cupin, expansin and aquaporin genes involved in cell wall synthesis and organization showed conserved downregulation.

A number of transporter and kinase/phosphatase genes showed clear patterns of coordinated expression under abiotic and biotic stresses. All three of the genes coding for pleiotropic drug resistance (PDR) type ATP-binding cassette (ABC) transporter proteins, which were found to be induced by ABA, SA and jasmonate in rice (Moons, 2008) showed conserved upregulation. Reversible protein phosphorylation executed by kinases and phosphatases is a fundamental mechanism that facilitates the orchestration of some of the most sophisticated signaling pathways. A number of different kinds of kinases and phosphatases were found in the list of common DEGs out of which serine/threonine protein kinases and phosphatases showed high distinction between the two stresses as also found by the GO analysis (Supplemental Table S4). All five of the protein phosphatase 2C (PP2C) genes showed conserved upregulation which are key players in ABA signaling pathways (Fig. 6). Four of these PP2C genes were part of the significant genes found by both SVM and PLS-DA multi-class indicating that these genes show
distinct pattern of expression in different stress conditions and can be considered as some of the most important genes to study multiple stress response.

A number of transporter and peroxidase (POX) precursor genes showed clear patterns of difference in expression between abiotic and biotic stresses. For instance, 2 out of 3 major facilitator superfamily (MFS) antiporter genes showed non-conserved upregulation under abiotic stresses. As many as 23 POX precursor genes were part of common DEGs, out of which 13 (56%) showed non-conserved downregulation under abiotic stresses. Further, 9 and 12 of these 23 POX genes were part of SVM and PLS-DA multi-class significant features, respectively. A study on rice infected with blast fungus showed ten POX genes redundantly respond to multiple stresses (Sasaki et al., 2004). Our findings suggest that the functionalities of many of the POX genes are specific to biotic stresses and are promising candidates to decipher the cross-talk between stresses.

The domain family with most number of conserved upregulated genes was Zinc Finger (ZF) family including C2H2, C3H TFs, C3HC4 and ZIM domain containing members with 14 and 15 members out of 17 showing overexpression in abiotic and biotic stresses, respectively. All of the pentatricopeptide (PPR) domain genes (eleven) which play essential roles in RNA editing, organelle biogenesis (Yuan and Liu, 2012) and plant development by coordinating interaction between mitochondria and chloroplasts (Toda et al., 2012) showed conserved downregulation except LOC_Os07g36450 which showed conserved upregulation. Thus, this gene would be an important candidate to further explore and understand their specific role under stress conditions and determine what makes it different from other PPR genes. Another interesting gene family showing high distinction between the two stress categories was LTP (protease inhibitor/seed storage/lipid transfer protein) with 5 out of 9 members showing non-conserved downregulation in abiotic stresses. VQ domain containing proteins were recently found to interact with WRKY TFs (WRKY33) in Arabidopsis. Further, knockout or overexpression of VQ substantially altered defense response (Cheng et al., 2012). There are 5 VQ domain genes in common DEGs out of which 4 showed non-conserved biotic upregulation. Further, WRKY24 which is the rice ortholog of WRKY33 also showed non-conserved biotic upregulation. The striking contrast of these set of genes in their behavior between abiotic and biotic stresses suggests them as important candidates to explore multiple stress response.
A list of studies that over-expressed or suppressed ten of the common DEGs that significantly altered the stress response are provided in Table IV. Seven of these are TF genes (four genes code for WRKY family of transcription factors) and are part of significant features found by SVM. Overexpression of a gene (LOC_Os01g55940) coding for an indole-3-acetic acid amido synthetase, conferred broad-spectrum resistance to *M. grisea*, *X. oryzae pv oryzae* and *pv oryzicola* (Fu et al., 2011). Overexpression of two genes coding for NAC transcription factors enhanced tolerance to multiple abiotic stresses viz. drought and salinity (LOC_Os03g60080) as well as cold (LOC_Os11g03300) (Hu et al., 2006; Jeong et al., 2010). The expression of these three genes and LOC_Os07g40290, an auxin responsive gene co-expressed with 40 other common DEGs, was upregulated in both abiotic and abiotic stresses. Further, we compared the common DEGs against a recently released database of *Arabidopsis* loss-of-function mutants (Lloyd and Meinke, 2012) using orthologous IDs which revealed 138 orthologous mutant genes out of which 33 showed increased resistance or sensitivity to a variety of stresses (Supplemental Table S7). Two genes (LOC_Os06g44010 and LOC_Os12g16720) are common between the DEGs in Table IV and their orthologs with loss-of-function mutants in *Arabidopsis*. The first gene encodes a WRKY transcription factor which is upregulated under biotic stress and down regulated under abiotic stress. The second gene codes for a cytochrome P450 monooxygenase whose inactivation leads to Sekiguchi lesion (SL1) mutant rice (Fujiwara et al., 2010). This gene is part of conserved upregulated gene set. Experimental analysis involving overexpression or knockout of stress responsive genes described above have often been studied with respect to one or a few stresses. Genetically engineering rice plants with the top candidate genes identified in our study, singly or in combination would identify their role in conferring broad range resistance to multiple abiotic and biotic stresses.

CONCLUSION

Availability of large volumes of genome scale gene expression data and advanced computational techniques enabled us to dissect the complex nature of stress response and examine in-depth the overlap between abiotic and biotic stress responses. Plethora of novel insights reported in this work revealed the overarching roles of major stress regulatory molecules including phytohormones such as ABA and JA/ET, parent compounds of small metabolites like
shikimate, transcription factors like WRKY and MYB, and signaling genes like WAKs which are central to the fine-tuning of stress response pathways. Further, the expression patterns exhibited by these genes provided molecular basis to classify different stress conditions with high accuracy. The top regulatory and signaling genes identified in this study are likely to be involved in cross talk between biotic and abiotic stress responses and provide potential candidates crucial for development of a rice variety with broad range stress tolerance. Further, mechanistic insights gained in rice on multiple stress responses would provide anchor points to explore specific stress signaling pathways and orthologous genes in other cereal crops.

**METHODS**

**Selection of Stress Response Microarray Studies and Identification of Differentially Expressed Genes**

All of the microarray studies performed on Affymetrix Rice Genome Array and deposited at Gene Expression Omnibus (GEO) under the platform GPL2025 were manually searched to identify and categorize 13 stress conditions (7 abiotic and 6 biotic stresses) as shown in Supplemental Table S1. Two meta-analysis studies were performed combining abiotic and biotic stresses separately. Briefly, the raw intensity CEL files of the selected samples were downloaded from GEO and intensity values were extracted from the CEL files using the bioconductor package Affy in R (Gautier et al., 2004), quality checked using the package, ArrayQualityMetrics (Kauffmann et al., 2009) and the samples failing quality tests were removed.

The samples of each stress were normalized together using Robust Multichip Average (RMA) method (Irizarry et al., 2003). The probes were then matched to their loci based on annotation provided at ricechip.org (http://www.ricechip.org). Probes with no match or ambiguously matching multiple loci were discarded. The retained probes and their normalized intensity values were then loaded into oneChannelGUI environment to perform non-specific filtering of probes with relatively small signal distribution using Inter Quartile Range (IQR) filter at most stringent setting (0.5) and probes with very low intensity values (probes below threshold log2(50)=5.64 in ≥90% of arrays). Differentially expressed genes (DEGs) were identified using
Rank Product method (Breitling et al., 2004). We used the function RPadvance of the bioconductor package RankProd (Hong et al., 2006) which is specifically designed for meta-analysis by taking into consideration the different origins of samples. The number of permutation tests was set to 250. The function topGene with a PFP (Percentage of False Positives) cut-off value of \( \leq 0.01 \) was used to output differentially expressed genes. Among multiple probes matching the same locus, the probe ID with highest fold change was retained.

**Classification Methods**

We used a number of classification and machine learning techniques to assess the performance of identified common DEGs between abiotic and biotic stresses in classification of different stresses. We extracted the RMA normalized intensity values of the identified common DEGs between abiotic and biotic stresses from stress treated microarrays (126 Abio and 232 Bio arrays) and scale adjusted using mean-centering and dividing by the square root of standard deviation of each variable (pareto scaling) (Fig. S1). Pareto scaling was chosen as it keeps the data structure partially intact while reducing the relative importance of large values (van den Berg et al., 2006).

Principal Component Analysis (PCA) is a non-supervised (i.e. does not make use of class labels) dimensionality reduction procedure which performs an orthogonal transformation of the original variables into a set of linearly uncorrelated variables such that the largest variance between the classes is captured in the transformed variables also called as principal components (PCs) (Yeung and Ruzzo, 2001). The PCs are numbered in decreasing order and the top PC (PC1) captures the maximal variance between different classes. Partial least squares Discriminant Analysis (PLS-DA) is a supervised (i.e. makes use of class labels) projection method that separates groups by rotating the PCs such that a maximum separation among classes is obtained (Zhang et al., 2013).

SVM classifies binary training data by drawing a hyper-plane (linear or nonlinear based on type of kernel selected) that maximally separates the two categories (Furey et al., 2000). R-SVM performs this type of classification recursively using different feature subsets and selects the best performing features based on cross-validation error rates. Although SVM based on microarray data is widely used to classify and predict disease status in humans (Hedenfalk et al.,
2001) and identify important features (Zhang et al., 2006), only a few studies have used R-SVM to identify stress responsive genes in plants (Liang et al., 2011). We performed R-SVM classification using linear kernel with genes (features) in columns and samples in rows. We utilized LOOCV (leave one out cross validation) procedure to determine the accuracy of the classification in which features are randomly partitioned into training and test sets and the poorly performing features with higher CV error rate are recursively eliminated. Random Forest (RF) is a decision tree based algorithm that grows the branches of an ensemble of classification trees by selecting random subsets of features from bootstrap samples and makes class prediction based on majority vote of the ensemble. A number of characteristics of RF make it ideal for our data set including its use for multi-class problems, less affected by noise and does not overfit the training data (Diaz-Uriarte and Alvarez de Andres, 2006). The statistical packages and tools provided by R, WEKA (Frank et al., 2004) and Metaboanalyst (Xia et al., 2012) were utilized to implement different analytical procedures.

Functional Enrichment Analysis

Gene ontology analysis was carried out using the Singular Enrichment Analysis (SEA) tool offered by agriGO(Du et al., 2010) at default settings of Fisher t-test (p<0.05), False Discovery Rate (FDR) correction by Hochberg method and five minimum number of mapping entries against species specific pre-computed background reference. Metabolic pathway enrichment analysis was carried out using the tool Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 (Huang et al., 2009). The functional annotation tool DAVID (Database for Annotation, Visualization and Integrated Discovery) v6.7 performs enrichment analysis of various annotation resources including gene ontologies, protein domains and pathways using a modified Fisher exact test called EASE. Further, it clusters significant annotation terms using kappa statistics and fuzzy heuristic clustering based on the degree of common genes between two annotations and provides an enrichment score for each annotation cluster. Information on transcription factors (TFs) genes in rice was obtained from the database PlnTFDB (Perez-Rodriguez et al., 2010) and analyzed for enrichment of TF families. The microRNAs predicted to target stress responsive genes were obtained from plant microRNA database (Zhang et al., 2010)
Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure S1. Box plot and density distribution of common DEGs before and after pareto normalization.

Supplemental Figure S2. Top 15 genes found by RF based on mean decreasing accuracy. A, mean decrease accuracy of top 15 genes given by RF for classification of ten stress conditions. B, and C, Box plots of normalized intensities of two of the top 15 genes found by RF.

Supplemental Figure S3. Top GO terms of common DEGs.

Supplemental Table S1.A. Number of microarray studies and samples analyzed for different stress conditions. B. Description of samples in different studies.

Supplemental Table S2.A. Differentially expressed genes in abiotic stresses. B. Differentially expressed genes in biotic stresses.

Supplemental Table S3. List of common DEGs.

Supplemental Table S4. Top functional annotation clusters found by the tool DAVID.

Supplemental Table S5. Distribution of TF families in conserved and non-conserved DEGs.

Supplemental Table S6. List of microRNAs targeting common DEGs.

Supplemental Table S7. List of common DEGs with loss-of-function mutant Arabidopsis orthologs and their mutant phenotype.
LITERATURE CITED


IRRI WRS. in http://www.irri.org/science/ricestat/index.asp,


Peng Y, Bartley LE, Chen X, Dardick C, Chern M, Ruan R, Canlas PE, Ronald PC (2008) OsWRKY62 is a negative regulator of basal and Xa21-mediated defense against Xanthomonas oryzae pv. oryzae in rice. Mol Plant 1: 446-458


Yanhui C, Xiaoyuan Y, Kun H, Meihua L, Jigang L, Zhaofeng G, Zhiqiang L, Yunfei Z, Xiaoxiao W,


**Figure Legends**

**Figure 1.** Comparison of differentially expressed genes (DEGs) under abiotic and biotic stress responses. A, Two-way Venn diagram showing the common DEGs between abiotic and biotic stresses. B, Number of up and downregulated DEGs in all of identified abiotic and biotic stresses. C, Four-way Venn diagram showing number of genes with conserved and non-conserved expression pattern.

**Figure 2.** Three dimensional plots of two-class classification of abiotic and biotic stresses. A and B, 3D plots based on top three components by PCA and PLS-DA, respectively using 1377 common DEGs. C and D, 3D plots based on top three components by PCA and PLS-DA, respectively using top 540 genes ranked by SVM return 100% accuracy of classification. The axes of B are rotated 90° which shows the best possible separation of two groups.

**Figure 3.** Classification error rates of different subsets of common DEGs upon 10-fold Cross Validation (CV) using R-SVM. Error rate using all of 1377 or 540 common DEGs was 0% (100% accuracy of classification) and 0.1% (99% of accuracy) using 220 genes and 0.5% (95% accuracy using 88 genes).

**Figure 4.** Multi-class classification of ten stress conditions by PLS-DA. All five abiotic stresses are circled by a red oval and all five biotic stresses by a green oval. A, Two-D plot between PLS-DA components 1 (14.9%) and 3 (8.1). B, Three-D plot between PLS-DA components 1 (14.9%), 2 (28.9%) and 3 (8.1%).

**Figure 5.** Co-expression network of common DEGs. The edges with r ≥ 0.95 are shown in red and r ≤ -0.9 are show in green. Nodes with >75 edges are shown in blue and >25 are shown in grey. The edges of NF-YC TF are shown in blue.

**Figure 6.** Visual representation of different gene families and functional categories based on expression between abiotic and biotic stresses. For each annotation, the stacked bars represent upregulated genes (total) scaled to 100% and downregulated genes scaled to -100%.
Table I. Classification of multiple stresses using Random Forest method

The overall Out-Of-Box (OOB) error rate was 0.0087

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Abio-Cold</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Abio-Drought</td>
<td>0</td>
<td>46</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Abio-Metal</td>
<td>0</td>
<td>0</td>
<td>29</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Abio-Nutrient</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Abio-Salt</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>26</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.037</td>
</tr>
<tr>
<td>Bio-Bacteria</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>166</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Bio-Fungus</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>23</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.08</td>
</tr>
<tr>
<td>Bio-Insect</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Bio-Nematode</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Bio-Weed</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Table II. Top 10 genes with highest VIP (Variable Importance in Projection) score in multi-class classification by PLS-DA

<table>
<thead>
<tr>
<th>MSU ID</th>
<th>Annotation</th>
<th>PLS-DA Multiple Stress (VIP comp.1)</th>
<th>PLS-DA AbioVsBio (VIP comp.1)</th>
<th>RF top 100 (MeanDecreaseAccuracy)</th>
<th>PLS-DA two-class (on SVM 540)</th>
<th>SVM Sig.540 (Freq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOC_Os01g09620</td>
<td>zinc finger/CCCH transcription factor</td>
<td>2.899</td>
<td>2.1512</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LOC_Os11g11970</td>
<td>expressed protein</td>
<td>2.8141</td>
<td>2.1813</td>
<td>0.003567</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LOC_Os11g26780</td>
<td>Dehydrin</td>
<td>2.8021</td>
<td>1.6548</td>
<td>0.002054</td>
<td>2.5959</td>
<td>35.8</td>
</tr>
<tr>
<td>LOC_Os07g48020</td>
<td>peroxidase</td>
<td>2.7902</td>
<td>1.8946</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LOC_Os06g4990</td>
<td>xylanase inhibitor protein 1</td>
<td>2.7427</td>
<td>2.2645</td>
<td>0</td>
<td>3.593</td>
<td>35.8</td>
</tr>
<tr>
<td>LOC_Os11g32800</td>
<td>expressed protein</td>
<td>2.718</td>
<td>1.6966</td>
<td>0</td>
<td>2.6919</td>
<td>35.8</td>
</tr>
<tr>
<td>LOC_Os06g48300</td>
<td>protein phosphatase 2C</td>
<td>2.6555</td>
<td>0</td>
<td>0</td>
<td>2.2334</td>
<td>35.8</td>
</tr>
<tr>
<td>LOC_Os10g40040</td>
<td>expressed protein</td>
<td>2.5979</td>
<td>0</td>
<td>0</td>
<td>2.3424</td>
<td>35.8</td>
</tr>
<tr>
<td>LOC_Os09g07350</td>
<td>fasciclin-like arabinoxylan protein 8</td>
<td>2.5059</td>
<td>0</td>
<td>0</td>
<td>1.9348</td>
<td>35.8</td>
</tr>
<tr>
<td>LOC_Os05g06920</td>
<td>RelA-SpoT like protein 8H4</td>
<td>2.5036</td>
<td>2.8245</td>
<td>0.004242</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Table III. Expression status of various hormone related genes in the common DEGs

* Number of orthologs of Arabidopsis plant hormone database genes are shown in brackets

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Abscisic acid</td>
<td>12 (9)</td>
<td>-</td>
<td>9 (3)</td>
<td>-</td>
<td>3 (1)</td>
</tr>
<tr>
<td>Auxin</td>
<td>21 (6)</td>
<td>5 (1)</td>
<td>11 (3)</td>
<td>3 (1)</td>
<td>2 (1)</td>
</tr>
<tr>
<td>Brassinosteroid</td>
<td>11 (5)</td>
<td>-</td>
<td>7 (4)</td>
<td>-</td>
<td>4 (1)</td>
</tr>
<tr>
<td>Cytokinin</td>
<td>4 (2)</td>
<td>4 (2)</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ethylene</td>
<td>10 (5)</td>
<td>3 (2)</td>
<td>1 (1)</td>
<td>-</td>
<td>6 (2)</td>
</tr>
<tr>
<td>Gibberellins</td>
<td>8 (3)</td>
<td>2 (1)</td>
<td>5 (2)</td>
<td>1 (0)</td>
<td>-</td>
</tr>
<tr>
<td>Jasmonic acid</td>
<td>7 (3)</td>
<td>4 (1)</td>
<td>-</td>
<td>1 (1)</td>
<td>2 (1)</td>
</tr>
<tr>
<td>Salicylic acid</td>
<td>5 (3)</td>
<td>-</td>
<td>4 (2)</td>
<td>-</td>
<td>1 (1)</td>
</tr>
<tr>
<td>MSU ID</td>
<td>Annotation</td>
<td>Phenotype</td>
<td>Reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>-------------------------------------------------</td>
<td>------------------------------------</td>
<td>-----------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOC_Os02g08440</td>
<td>WRKY71</td>
<td>Enhanced defense response</td>
<td>(Liu et al., 2007)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOC_Os03g60080</td>
<td>NAC domain-containing protein 67</td>
<td>Increased drought and salt tolerance</td>
<td>(Hu et al., 2006)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOC_Os05g25770</td>
<td>WRKY45</td>
<td>Increased susceptibility to bacteria</td>
<td>(Tao et al., 2009)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOC_Os06g44010</td>
<td>WRKY28</td>
<td>Enhanced disease resistance</td>
<td>(Peng et al., 2010)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOC_Os07g40290</td>
<td>OsGH3.8 - Probable indole-3-acetic acid-amido synthetase</td>
<td>Enhanced disease resistance</td>
<td>(Ding et al., 2008)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOC_Os08g06280</td>
<td>LSD1 zinc finger domain containing protein</td>
<td>Increased resistance to blast fungus</td>
<td>(Wang et al., 2005)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOC_Os09g25070</td>
<td>WRKY62</td>
<td>Increased bacterial susceptibility</td>
<td>(Peng et al., 2008)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOC_Os11g03300</td>
<td>NAC domain transcription factor</td>
<td>Increased drought tolerance and yield</td>
<td>(Jeong et al., 2010)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOC_Os12g16720</td>
<td>cytochrome P450 71A1</td>
<td>Enhanced fungal resistance*</td>
<td>(Fujiwara et al., 2010)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* - suppression of gene expression by knockout; N/A - Not Applicable

Table IV. List of common DEGs which showed alteration in stress response upon over-expression/suppression
Recursive SVM classification

Error Rate vs. Number of variables (levels)