Supplementary Information for

Predicting transcriptional responses to cold stress across plant species

Xiaoxi Meng, Zhikai Liang, Xiuru Dai, Yang Zhang, Samira Mahboub, Daniel W. Ngu, Rebecca L. Roston, and James C. Schnable

James C. Schnable.
E-mail: schnableunl.edu

This PDF file includes:

Figs. S1 to S12
Legend for Dataset S1
SI References

Other supplementary materials for this manuscript include the following:

Dataset S1

Supplementary Figures

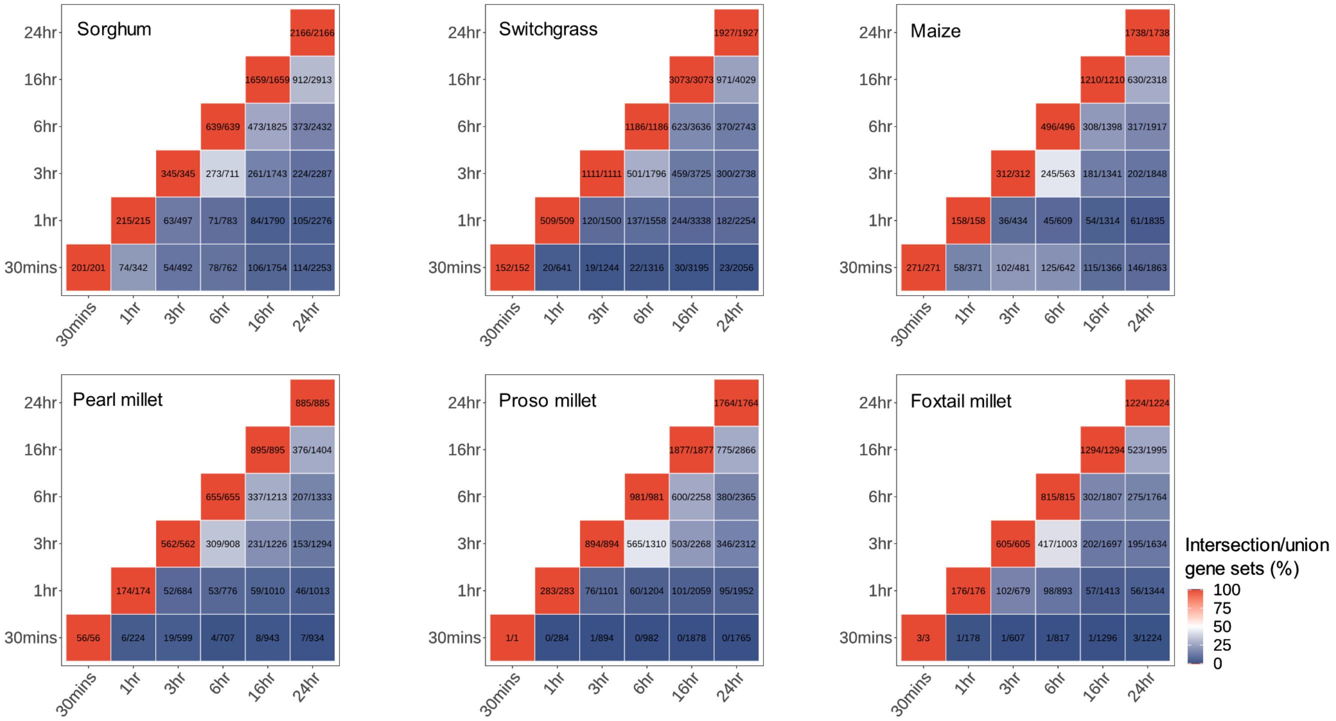


Fig. S1. Overlap of cold-responsive genes among time points in each species. Abundance of intersecting differentially expressed genes as a percent of the union of differentially expressed genes between pairs of time points in each grass species analyzed in this study. In each cell the numerator indicates the intersection of the sets of differentially expressed genes identified at the two time points and the indicates the union of the same two sets of genes.

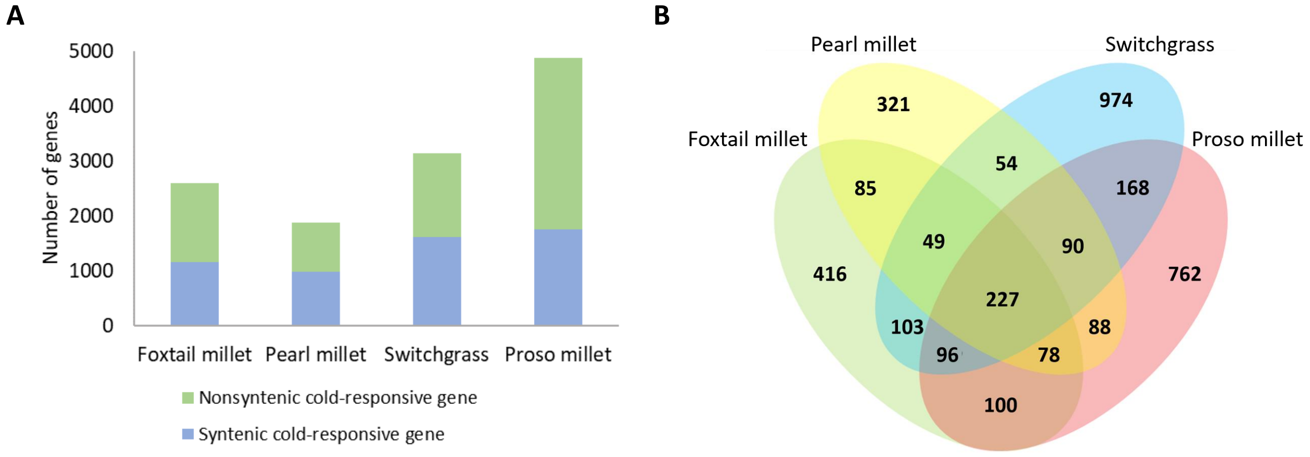


Fig. S2. Conserved cold-responsive genes across foxtail millet, pearl millet, switchgrass, and proso millet. A. Proportions of syntenic orthologous genes among cold-responsive genes; B. Overlapping cold-responsive syntenic orthologs among the four species.

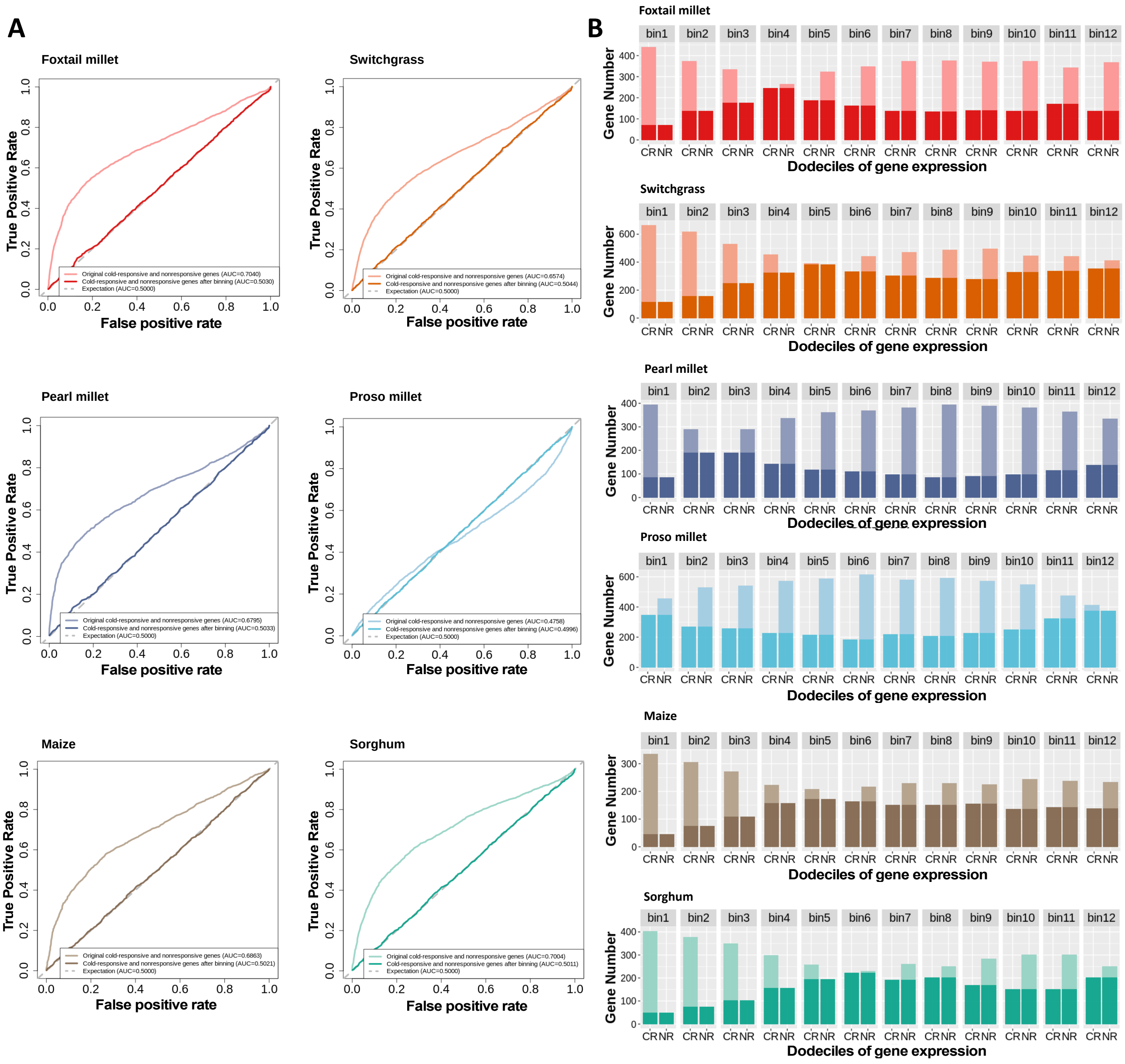


Fig. S3. Baseline expression controls. A. Accuracy of genes being scored as cold-responsive genes solely based on average FPKM values before and after baseline expression control; B. Distribution of average FPKM values of cold-responsive genes (CR) and nonresponsive genes (NR), and training sets resampled from genes in dodeciles with balanced gene expression levels (darker color).

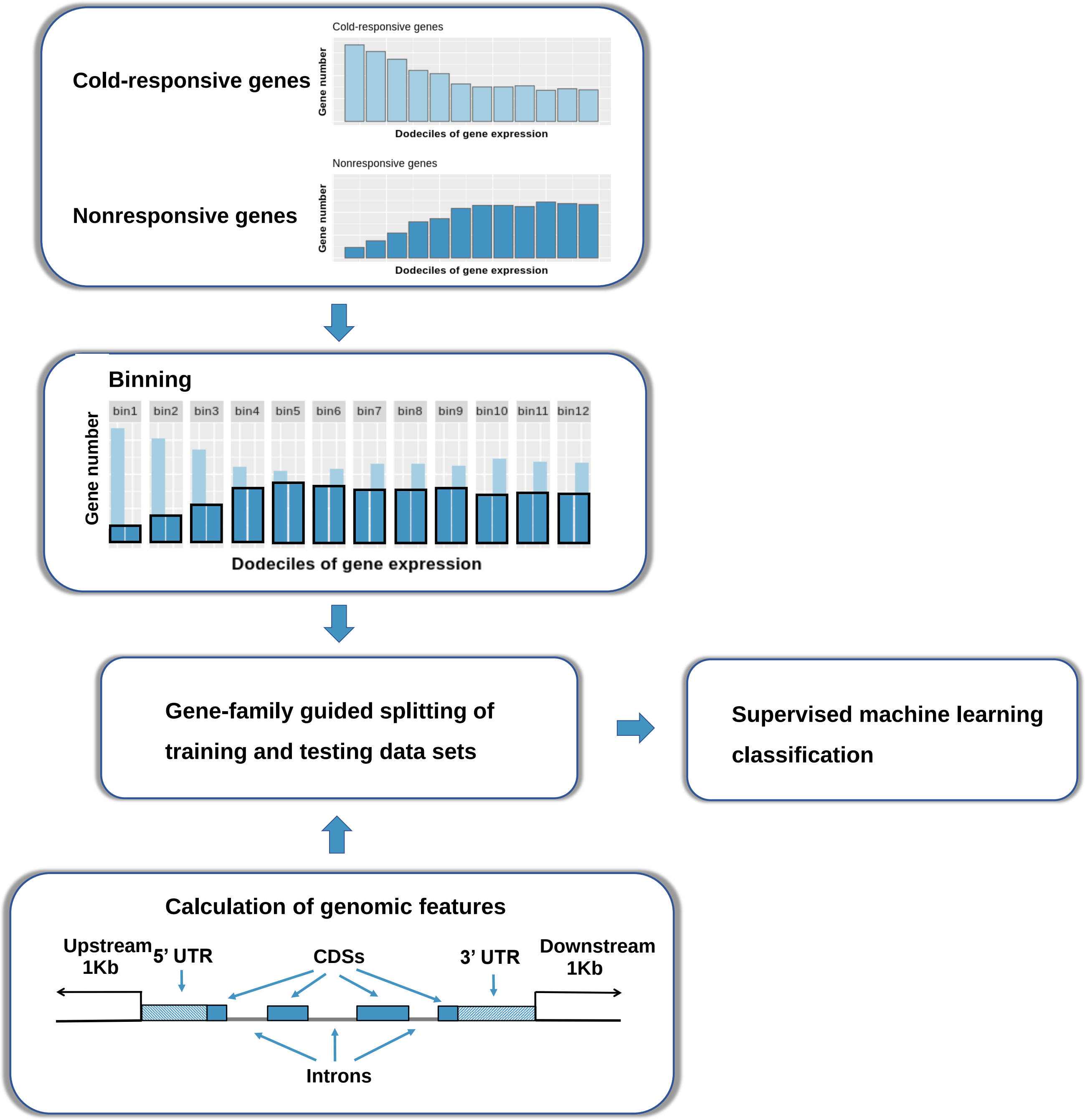


Fig. S4. Workflow of the supervised machine classification model for predicting cold-responsive genes. For within species predictions, gene-family guided subsampling and splitting consisted of first subsampling each gene family present in the species and then dividing into training/validation and testing data. For cross-species predictions, gene-family guided subsampling and splitting consisted of first dividing gene families into training/validation and testing data and then subsampling one gene per family per species.

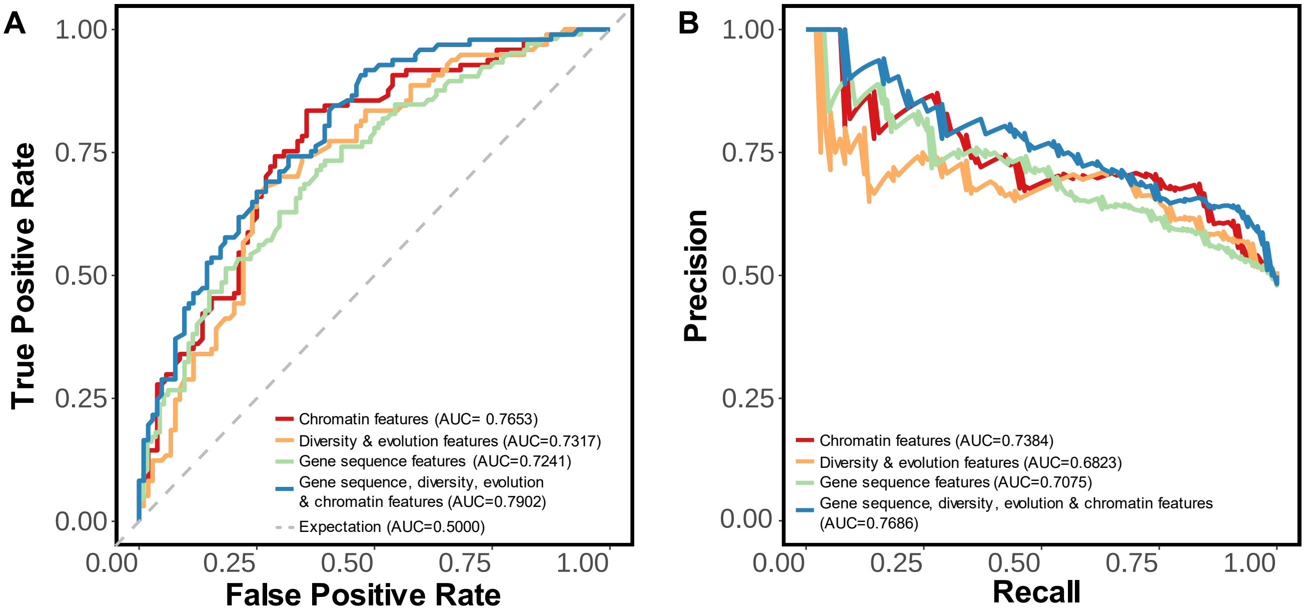


Fig. S5. Cold-responsive gene predictions in maize using different subsets of features A. Receiver operating characteristic (ROC) curves shows the classification on holdout test data; B. Precision-recall (PR) curves shows the classification on holdout test data.

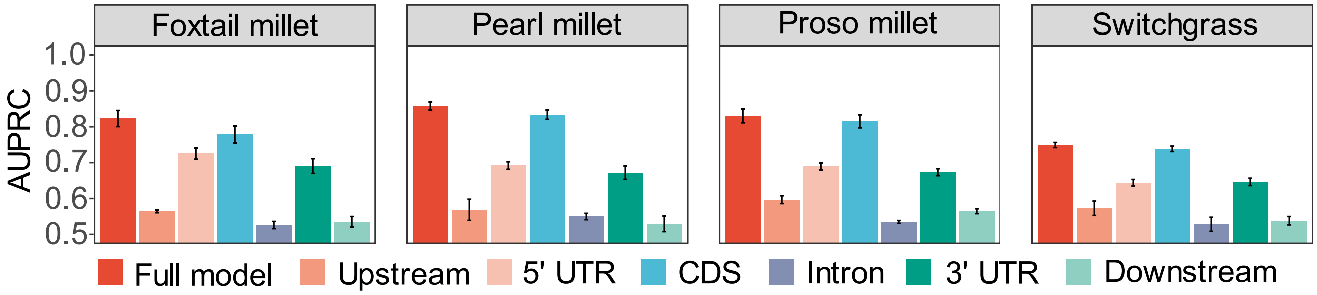


Fig. S6. AUPRCs of supervised machine learning models for Paniceae grass species based on gene sequence features. Bar plot showing AUPRCs achieved by the full gene sequence models and single feature group models for foxtail millet, pearl millet, switchgrass and proso millet. Standard error (se) was calculated from five independent predictions.

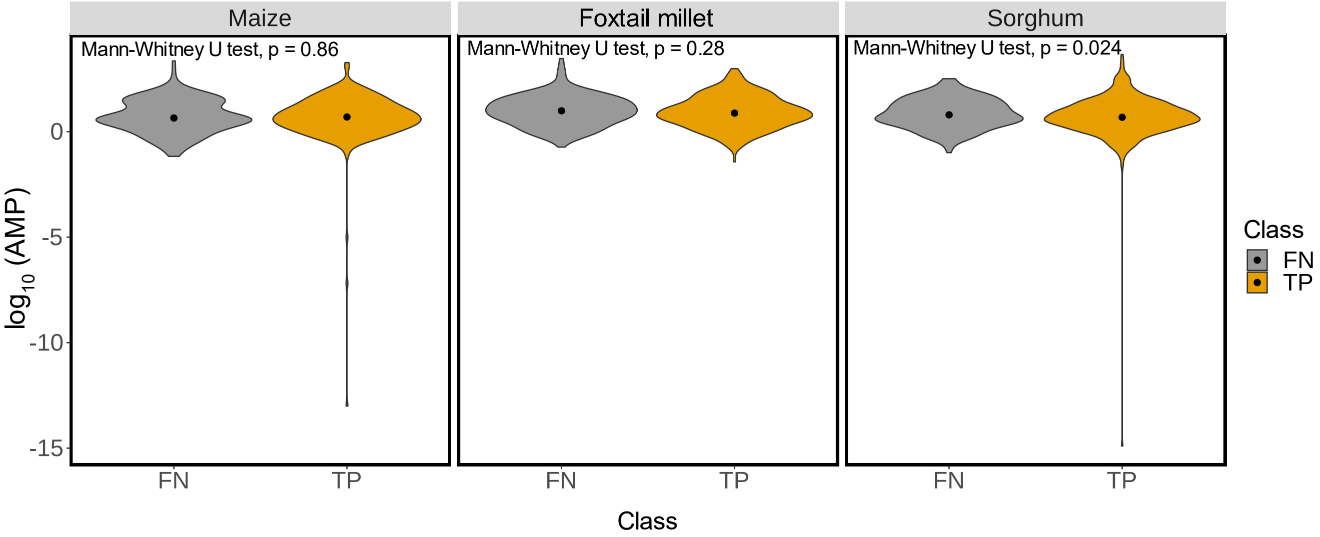


Fig. S7. Model performance evaluations on predicting cold-responsive genes and circadian genes. The trained machine learning model in each species (Maize, Sorghum and Foxtail millet) using gene sequence features was applied on predicting holdout test data. From prediction results, genes in holdout test data were split into true positive (TP) and false negative (FN) sets. Amplitudes of unique genes were considered together in one group and the black dot indicate the median value of amplitudes in each group. Mann-whitney U test was applied on comparing raw amplitudes between groups in each species. AMP represents amplitude.

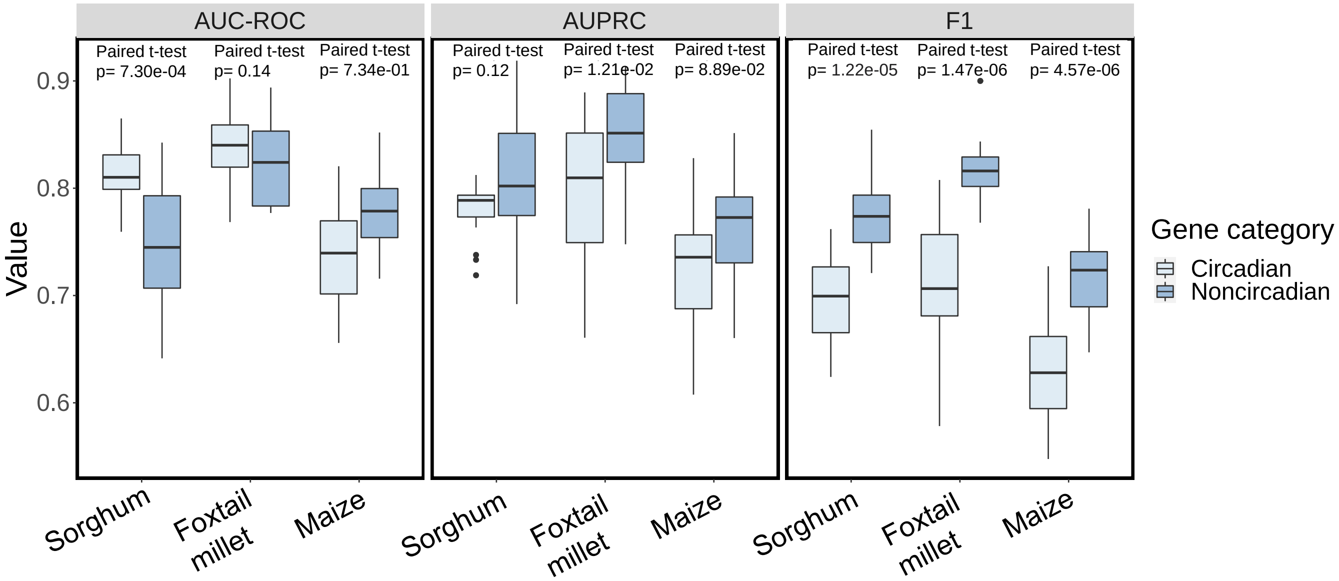


Fig. S8. Model performance evaluations on predicting circadian genes. The trained machine learning model in each species (Maize, Sorghum and Foxtail millet) using gene sequence features was applied on predicting diurnal cycling genes. AUC-ROC, AUPRC and F1 values were calculated on 10% holdout test data based on 20 prediction models per species. Paired t-test was used to evaluate statistical significance between samples.

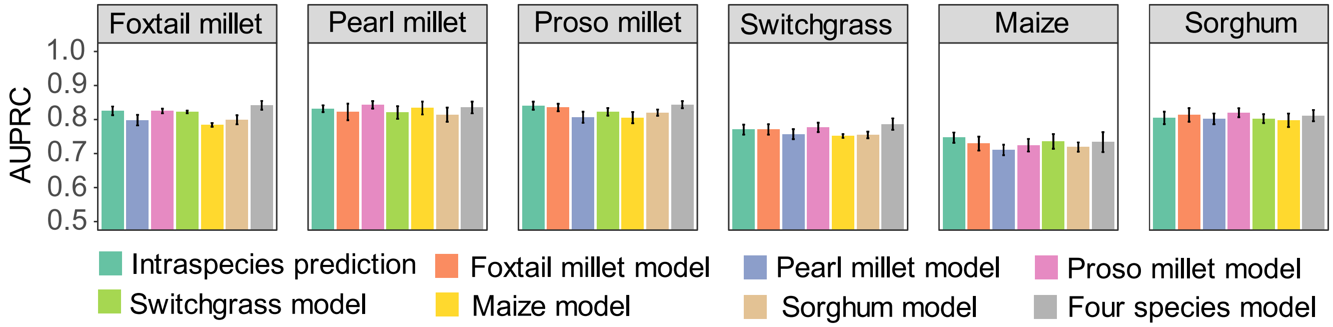


Fig. S9. AUPRC of models trained for cross-species prediction. Areas under Precision-Recall Curves (AUPRC) show the classification on holdout test data in machine learning models constructed in different species. Standard error (se) was calculated from five independent predictions. All predictions shown here, including intraspecies predictions, were made using the cross-species prediction framework for partitioning hold out test data (see methods).

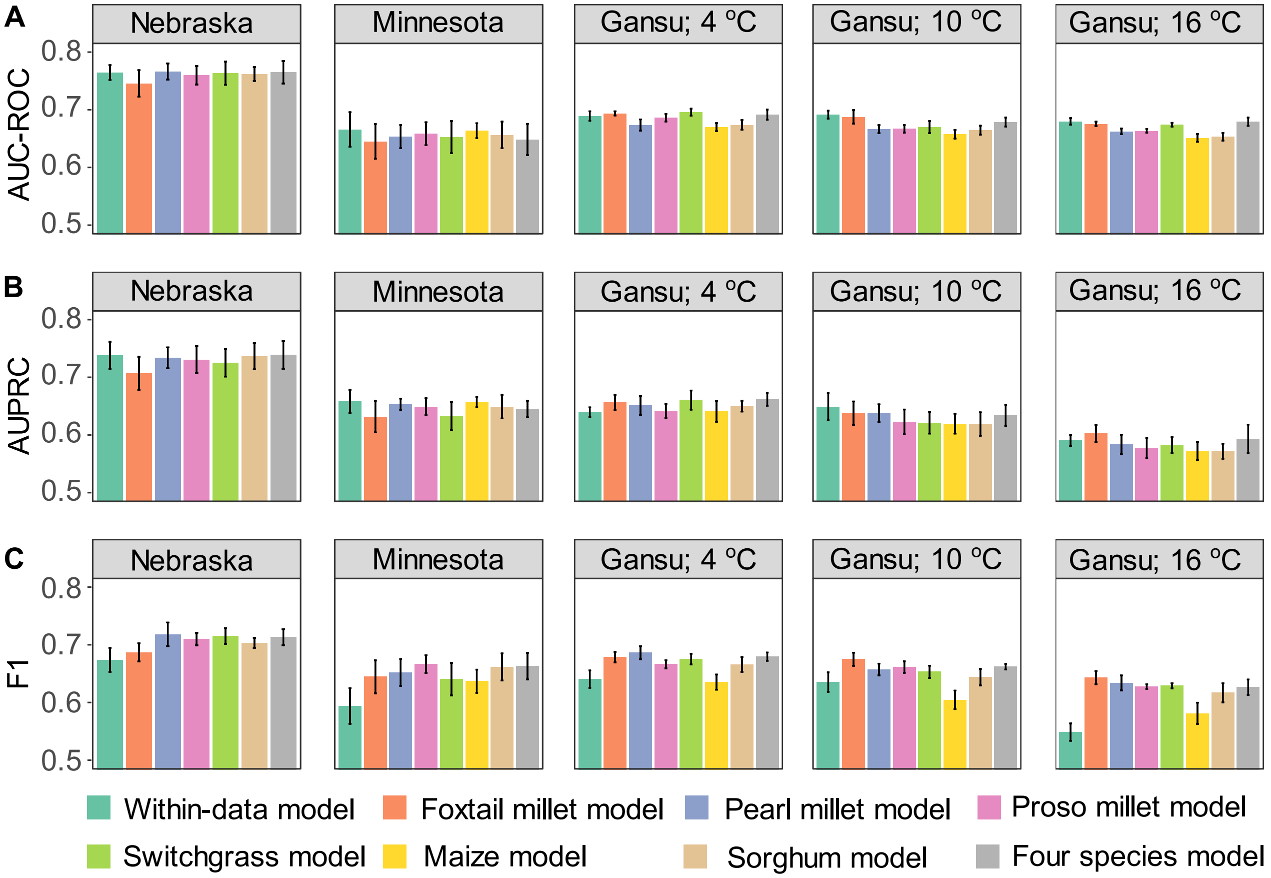


Fig. S10. Comparisons of model performance in predicting cold-responsive genes in maize identified using RNA seq data collected by different research groups. A. Areas under the receiver operating characteristic curves (AUC-ROCs) for predicting cold-responsive genes identified in different maize data sets. Standard error (se) was calculated from five independent predictions. Minnesota indicates data generated by Liang *et al.*, 2020 (1). Gansu (China) indicates data generated by Li *et al.*, 2020 (2) with different cold temperatures as labeled. Nebraska and "Maize model" indicate the same data employed in this study; B. Areas under Precision-Recall Curves (AUPRC) show performance of predictions on cold-responsive genes; C. F1 scores show performance of predictions on cold-responsive genes.

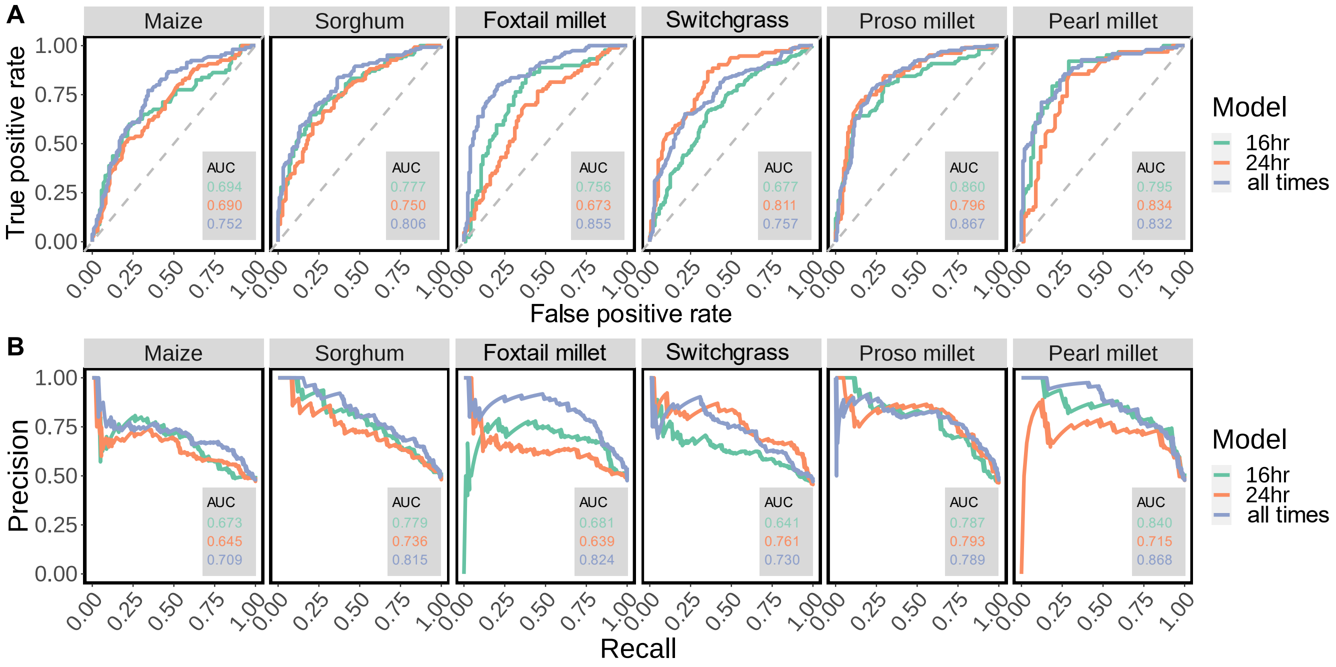
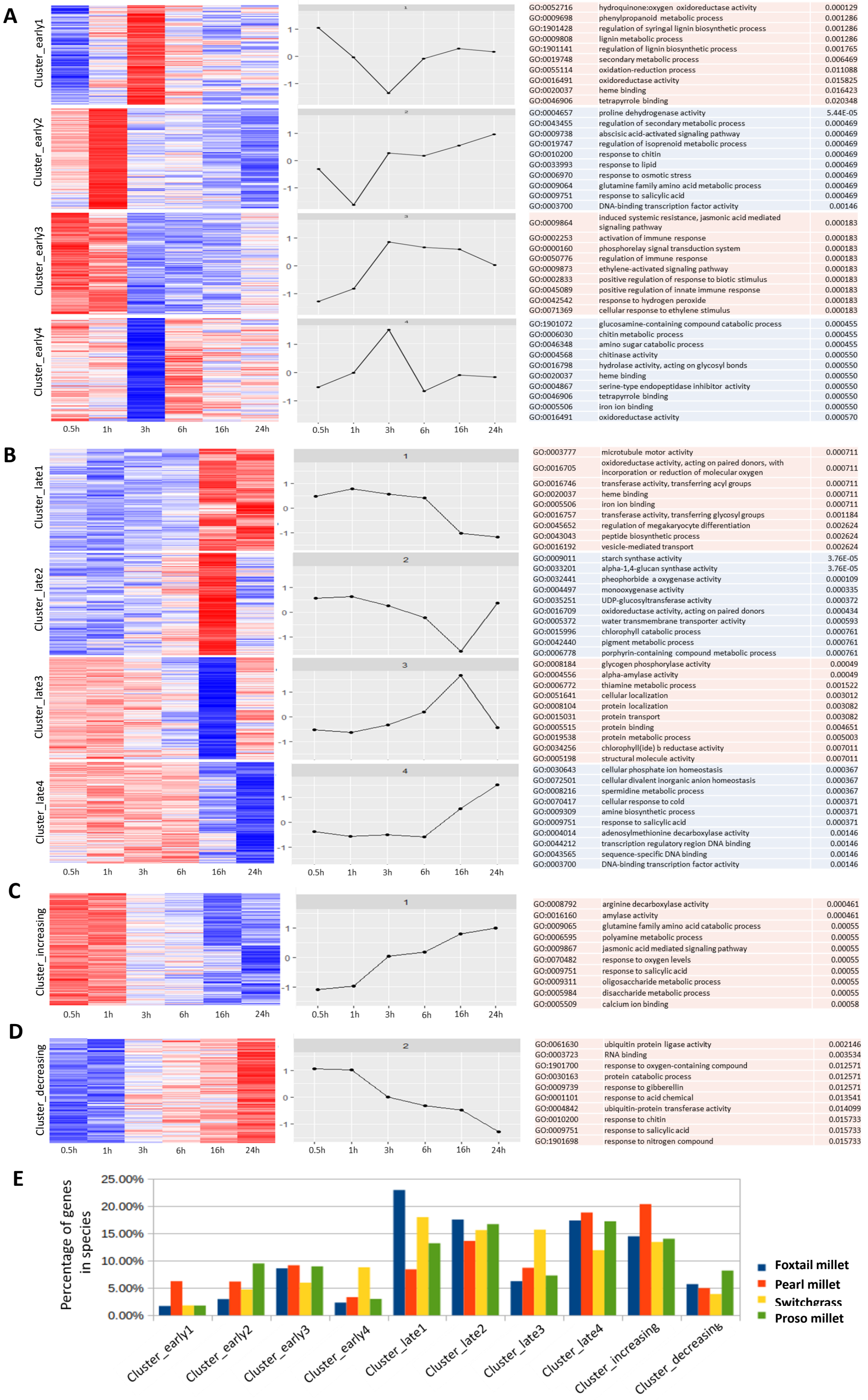


Fig. S11. Comparisons of model performance on predicting cold-repsonsive genes defined from single and multiple time points in six grass species. A. Receiver operating characteristic (ROC) curves show classifications on holdout test data. Data collected from 16 hr and 24 hr represented single time point data. All times are the same as we defined cold-responsive genes in the study. Values of AUC were indicated within each species with same colors for corresponding models; B. Precision-recall (PR) curves show the classification on holdout test data.



GO:0052716	hydroquinone:oxygen oxidoreductase activity	0.000129
GO:0009698	phenylpropanoid metabolic process	0.001286
GO:1901428	regulation of syringal lignin biosynthetic process	0.001286
GO:0009808	lignin metabolic process	0.001286
GO:1901141	regulation of lignin biosynthetic process	0.001765
GO:0019748	secondary metabolic process	0.006469
GO:0055114	oxidation-reduction process	0.011088
GO:0016491	oxidoreductase activity	0.015825
GO:0020037	heme binding	0.016423
GO:0046906	tetrapyrrole binding	0.020348
GO:0004657	proline dehydrogenase activity	5.44E-05
GO:0043455	regulation of secondary metabolic process	0.000469
GO:0009738	abscisic acid-activated signaling pathway	0.000469
GO:0019747	regulation of isoprenoid metabolic process	0.000469
GO:0010200	response to chitin	0.000469
GO:0033993	response to lipid	0.000469
GO:0006970	response to osmotic stress	0.000469
GO:0009064	glutamine family amino acid metabolic process	0.000469
GO:0009751	response to salicylic acid	0.000469
GO:0003700	DNA-binding transcription factor activity	0.00146
GO:0009864	induced systemic resistance, jasmonic acid mediated signaling pathway	0.000183
GO:0002253	activation of immune response	0.000183
GO:0000160	phosphorelay signal transduction system	0.000183
GO:0050776	regulation of immune response	0.000183
GO:0009873	ethylene-activated signaling pathway	0.000183
GO:0002833	positive regulation of response to biotic stimulus	0.000183
GO:0045089	positive regulation of innate immune response	0.000183
GO:0042542	response to hydrogen peroxide	0.000183
GO:0071369	cellular response to ethylene stimulus	0.000183
GO:1901072	glucosamine-containing compound catabolic process	0.000455
GO:0006030	chitin metabolic process	0.000455
GO:0046348	amino sugar catabolic process	0.000455
GO:0004568	chitinase activity	0.000550
GO:0016798	hydrolase activity, acting on glycosyl bonds	0.000550
GO:0020037	heme binding	0.000550
GO:0004867	serine-type endopeptidase inhibitor activity	0.000550
GO:0046906	tetrapyrrole binding	0.000550
GO:0005506	iron ion binding	0.000550
GO:0016491	oxidoreductase activity	0.000570

GO:0003777	microtubule motor activity	0.000711
GO:0016705	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	0.000711
GO:0016746	transferase activity, transferring acyl groups	0.000711
GO:0020037	heme binding	0.000711
GO:0005506	iron ion binding	0.000711
GO:0016757	transferase activity, transferring glycosyl groups	0.001184
GO:0045652	regulation of megakaryocyte differentiation	0.002624
GO:0043043	peptide biosynthetic process	0.002624
GO:0016192	vesicle-mediated transport	0.002624
GO:0009011	starch synthase activity	3.76E-05
GO:0033201	alpha-1,4-glucan synthase activity	3.76E-05
GO:0032441	pheophorbide a oxygenase activity	0.000109
GO:0004497	monooxygenase activity	0.000335
GO:0035251	UDP-glucosyltransferase activity	0.000372
GO:0016709	oxidoreductase activity, acting on paired donors	0.000434
GO:0005372	water transmembrane transporter activity	0.000593
GO:0015996	chlorophyll catabolic process	0.000761
GO:0042440	pigment metabolic process	0.000761
GO:0006778	porphyrin-containing compound metabolic process	0.000761
GO:0008184	glycogen phosphorylase activity	0.00049
GO:0004556	alpha-amylase activity	0.00049
GO:0006772	thiamine metabolic process	0.001522
GO:0051641	cellular localization	0.003012
GO:0008104	protein localization	0.003082
GO:0015031	protein transport	0.003082
GO:0005515	protein binding	0.004651
GO:0019538	protein metabolic process	0.005003
GO:0034256	chlorophyll(ide) b reductase activity	0.007011
GO:0005198	structural molecule activity	0.007011
GO:0030643	cellular phosphate ion homeostasis	0.000367
GO:0072501	cellular divalent inorganic anion homeostasis	0.000367
GO:0008216	spermidine metabolic process	0.000367
GO:0070417	cellular response to cold	0.000371
GO:0009309	amine biosynthetic process	0.000371
GO:0009751	response to salicylic acid	0.000371
GO:0004014	adenosylmethionine decarboxylase activity	0.00146
GO:0044212	transcription regulatory region DNA binding	0.00146
GO:0043565	sequence-specific DNA binding	0.00146
GO:0003700	DNA-binding transcription factor activity	0.00146

GO:0008792	arginine decarboxylase activity	0.000461
GO:0016160	amylase activity	0.000461
GO:0009065	glutamine family amino acid catabolic process	0.00055
GO:0006595	polyamine metabolic process	0.00055
GO:0009867	jasmonic acid mediated signaling pathway	0.00055
GO:0070482	response to oxygen levels	0.00055
GO:0009751	response to salicylic acid	0.00055
GO:0009311	oligosaccharide metabolic process	0.00055
GO:0005984	disaccharide metabolic process	0.00055
GO:0005509	calcium ion binding	0.00058

GO:0061630	ubiquitin protein ligase activity	0.002146
GO:0003723	RNA binding	0.003534
GO:1901700	response to oxygen-containing compound	0.012571
GO:0030163	protein catabolic process	0.012571
GO:0009739	response to gibberellin	0.012571
GO:0001101	response to acid chemical	0.013541
GO:0004842	ubiquitin-protein transferase activity	0.014099
GO:0010200	response to chitin	0.015733
GO:0009751	response to salicylic acid	0.015733
GO:1901698	response to nitrogen compound	0.015733

Fig. S12. Gene expression clusters analysis A-D. Cold-responsive genes from foxtail millet, pearl millet, switchgrass, and proso millet were analyzed using k-means clustering. This process identified eight major groups, as shown in heat map and graphical format, based on patterns of gene expression at different time points. (A) Clusters containing genes of early transcriptional responses to cold (30 min to 3 h); (B) clusters with late responded genes to cold (responded after 6h); (C and D) genes with continuously increasing or decreasing transcriptional levels within 24hrs. Enriched GO terms within clusters were shown in the last column. E. Percentage of genes of each of the four species distributed in clusters.

SI Dataset S1 ()

Dataset S1, Tab1 Number of cold-responsive genes identified in the four grass species

Dataset S1, Tab2 Performance metrics on predicting maize cold responsive genes by models with or without considering evolutionary relatedness and baseline expression.

Dataset S1, Tab3 Gene sequence features, chromatin features, and diversity/evolutionary features used in supervised machine learning classification.

Dataset S1, Tab4 The top 20 random forest feature importance presented by Mean Decrease Accuracy for each intraspecies prediction of transcriptional responses to cold stress.

Dataset S1, Tab5 Model performance on predicting cold responsive gene sets identified across species.

Dataset S1, Tab6 Model performance on predicting maize cold responsive gene sets identified by different experiments.

Dataset S1, Tab7 Syntenic gene list among *S. bicolor*, *S. italica*, *P. glaucum*, *P. miliaceum*, and *P. virgatum.*

References

1. Z Liang, et al., Genetic and epigenetic contributions to variation in transposable element expression responses to abiotic stress in maize. *bioRxiv* (2020).
2. Y Li, et al., Transcriptomic analysis revealed the common and divergent responses of maize seedling leaves to cold and heat stresses. *Genes* **11**, 881 (2020).