PNAS RESEARCH ARTICLE | PLANT BIOLOGY

Cross-species predictive modeling reveals conserved drought responses between maize and sorghum

Jeremy Pardo[a,b,c], Ching Man Wai[a,b], Maxwell Harman[a], Annie Nguyen[a,b], Karl A. Kremling[d,e], Maria Cinta Romay[d,e], Nicholas Lepak[f], Taryn L. Bauerle[e], Edward S. Buckler[d,e,f], Addie M. Thompson[b,g], and Robert VanBuren[a,b,1]

Edited by Donald Ort, University of Illinois at Urbana Champaign, Urbana, IL; received October 3, 2022; accepted January 30, 2023

Drought tolerance is a highly complex trait controlled by numerous interconnected pathways with substantial variation within and across plant species. This complexity makes it difficult to distill individual genetic loci underlying tolerance, and to identify core or conserved drought-responsive pathways. Here, we collected drought physiology and gene expression datasets across diverse genotypes of the C4 cereals sorghum and maize and searched for signatures defining water-deficit responses. Differential gene expression identified few overlapping drought-associated genes across sorghum genotypes, but using a predictive modeling approach, we found a shared core drought response across development, genotype, and stress severity. Our model had similar robustness when applied to datasets in maize, reflecting a conserved drought response between sorghum and maize. The top predictors are enriched in functions associated with various abiotic stress-responsive pathways as well as core cellular functions. These conserved drought response genes were less likely to contain deleterious mutations than other gene sets, suggesting that core drought-responsive genes are under evolutionary and functional constraints. Our findings support a broad evolutionary conservation of drought responses in C4 grasses regardless of innate stress tolerance, which could have important implications for developing climate resilient cereals.

predictive modeling | C4 grasses | maize | drought | transfer learning

Drought is responsible for billions of US dollars in losses each year, and the impacts of drought are most severe in developing regions of the world where food security is already low (1). Water deficit elicits hundreds to thousands of interconnected molecular pathways in plants, and drought tolerance represents a complex, emergent phenotype that is challenging to breed for or separate into major genetic loci (2, 3). Drought is also a difficult stress to apply and quantify, and plant responses to physiologically relevant drought events in the field are often different from those detected under controlled experiments in growth chamber or greenhouse settings (4, 5). These compounding issues represent major challenges for studying drought stress, but they also present an opportunity to leverage systems level and predictive modeling-based approaches to understand complex traits in plants.

C4 grasses dominate natural and agricultural settings, and they have evolved a unique set of adaptations that enable an emergent resilience to drought and other abiotic stresses (6). *Sorghum bicolor* (sorghum) is one of the most stress tolerant and highly productive C4 cereals, and it is an important agricultural commodity grown globally for grain, sugar, and biomass. Sorghum was domesticated in the semi-arid Sudanese savannah of northeast Africa around 4,000 BCE (7), and subsequently spread westward across the African steppe and throughout the Indian subcontinent and China. The broad geographic and climatic regions where sorghum was historically cultivated has led to significant diversity and local adaptation among cultivars. While sorghum is generally regarded as a drought tolerant crop, there remains considerable variation for abiotic stress tolerance among different sorghum accessions.

Drought tolerance is a highly complex trait in sorghum and numerous developmental and morphophysiological traits have been correlated with tolerance. Sorghum cultivars are often classified as either preflowering or postflowering drought tolerant with tolerance at both developmental stages a relative rarity (8). Postflowering drought tolerance is related to stay-green traits that prevent premature senescence (9, 10). Preflowering drought is characterized by more varied responses, and reactive oxygen species scavenging, cuticular wax production, and flowering time regulation are important components of preflowering drought tolerance in sorghum (5, 10, 11). Numerous previous studies have examined the transcriptomic response of different genotypes to both preflowering and postflowering drought tolerance in sorghum (11–13). However, these studies focused on the response of two or a few genotypes, limiting their ability to identify conserved and divergent patterns of expression across the diversity of cultivated sorghum.

Significance

Drought is a complex and variable stress that is difficult to quantify and link to underlying mechanisms both within and across species. Here, we developed a predictive model to classify drought stress responses in sorghum and identify important features that are responsive to water deficit. Our model has high predictive accuracy across development, genotype, and stress severity, and the top features are enriched in genes related to classical stress responses and have functional and evolutionary conservation. We applied this sorghum-trained model to maize, and observed similar predictive accuracy of drought responses, supporting transfer learning across plant species. Our findings suggest there are deeply conserved drought responses across C4 grasses that are unrelated to tolerance.

Author contributions: J.P., A.M.T., and R.V. designed research; J.P., C.M.W., M.H., A.N., K.A.K., M.C.R., N.L., and R.V. performed research; C.M.W., T.L.B., E.S.B., and A.M.T. contributed new reagents/analytic tools; J.P., M.H., A.N., and R.V. analyzed data; and J.P. and R.V. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: bob.vanburen@gmail.com.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2216894120/-/DCSupplemental.

Published February 27, 2023.

The broad genetic diversity of sorghum is captured by the sorghum association panel (SAP), which is composed of 400 temperate breeding lines as well as converted tropical lines that collectively represent the bulk of sorghum diversity (14). Association studies have identified genomic regions linked with drought response in sorghum (15) but unlike expression studies, it is challenging to link specific genes to underlying phenotypes. Here, we compared gene expression across the SAP during a natural drought event and leveraged these data to identify conserved and variable drought responses across sorghum lines.

We hypothesize that a core and deeply conserved drought response operates both within sorghum germplasm and across related species, reflecting ancestral adaptations of C4 grasses. Prior studies have observed commonalities in differentially expressed genes under drought stress across diverse angiosperms, but these studies were limited in sampling, species, or tissue breadth (16, 17). The progenitors of sorghum and *Zea mays* (maize) diverged 11.9 Mya, and maize and sorghum still share many similar morphological, biochemical, and genetic traits (18). However, sorghum is more drought tolerant than maize (19), creating an ideal comparative system. The drought responses of sorghum and maize have been compared using only one or a few genotypes (19, 20), but these studies are limited because they fail to account for the broad intraspecific variation present in both species.

Here, we compared interspecific variation and conservation of drought response between maize and sorghum as well as intraspecific variation in both species individually. We generated drought and well-watered expression data across 25 diverse sorghum genotypes and 27 diverse maize genotypes. We also leveraged additional new and public sorghum drought datasets (10–12, 21, 22) to develop a predictive model capable of classifying samples as drought responsive based on gene expression. We dissected the model to identify genes involved in drought response and applied our model to maize to elucidate evolutionary conserved patterns across both species.

Results

Climate-Relevant Drought Responses across Sorghum Accessions. Physiologically relevant drought stresses are difficult to simulate in controlled settings, and we sought to capture sorghum responses to a natural drought event in an agricultural setting. East Lansing, Michigan, experienced a period of below average precipitation corresponding to a mild drought event between June and early July 2020 (Fig. 1*A*). We collected physiological and RNA samples from 25 diverse sorghum genotypes during this natural stress event and 4 d later after a heavy rainfall event where plants recovered. We found that the sorghum plants had significantly low relative water content during the dry period compared with recovery, suggesting the plants were experiencing mild water-deficit stress (Fig. 1*B*). We also found that the sorghum plants had higher instantaneous nonphotochemical quenching, as measured by the photosynthesis parameter non-photochemical quenching (NPQ_t), during the dry period (Fig. 1*C*) (23). Nonphotochemical quenching increases under drought stress as a mechanism to dissipate excess light energy when photosynthesis is carbon limited as a result of stomatal closure (24). However, despite the drop in NPQ_t, we did not detect any differences in photosynthetic efficiency (ΦII) or linear electron flow, suggesting that the light reactions of photosynthesis were still proceeding at a high pace (Fig. 1*D*). Together, this suggests the sorghum plants were experiencing a very mild, and fully recoverable drought event.

We collected three replicates of RNA sequencing (RNAseq) data for each genotype at the drought and recovery timepoints to search for expression patterns corresponding to water-deficit

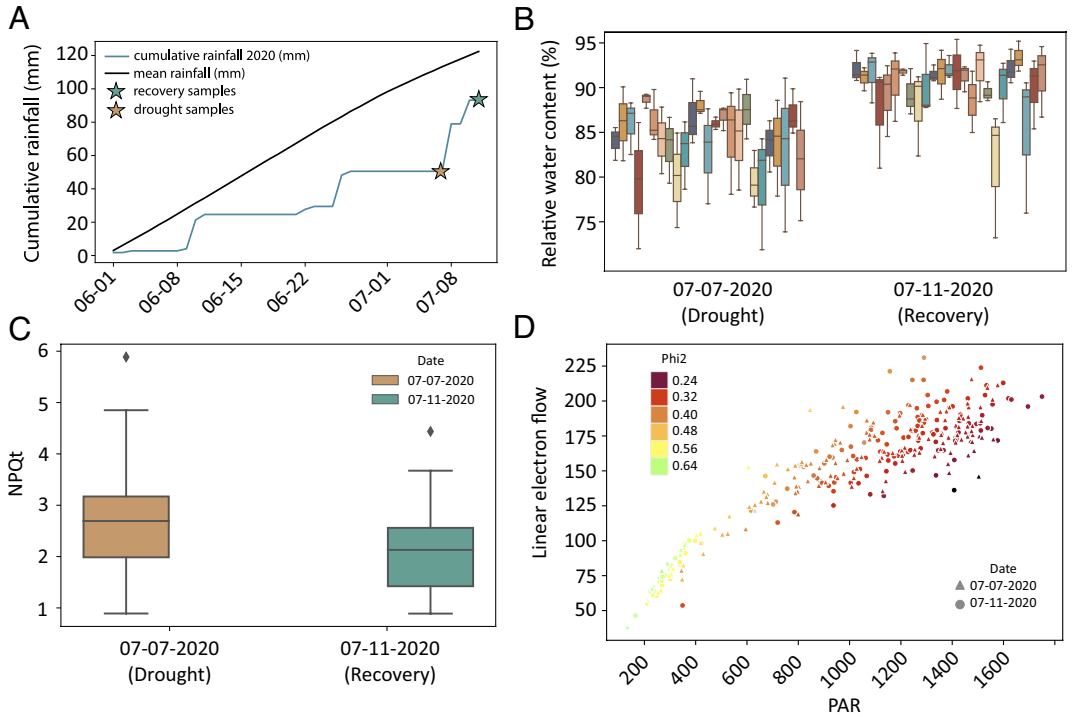

Fig. 1. Physiological response of diverse field grown sorghum genotypes across a natural drought stress event. (*A*) Cumulative growing season precipitation before and during the sampling period compared to the 30-y mean. The two sampling dates are labeled with stars. (*B*) Box plots of relative water content for each of the 25 genotypes on the two sampling dates. (*C*) Boxplot of NPQt on each sampling date. (*D*) Scatterplot showing linear electron flow as a function of photosynthetically active radiation (PAR). Points are colored by photosystem II efficiency (Phi2) with circles representing the recovery (November 7, 2020) sampling date and triangles representing the drought (July 7, 2020) sampling date.

responses in sorghum. Dimensionality reduction analysis clearly separates the RNAseq samples into two distinct groups of well-watered and drought along principal component 1 (Fig. 2*A*). Principal component 2 separates the samples of sorghum by genotype, but this pattern is not absolute (*SI Appendix*, Fig. S1). Across all genotypes, we identified 1,761 genes up-regulated under water deficit, and 2,317 genes down-regulated. Among up-regulated genes under drought, we found enrichment of gene ontology terms related to stress responses, including response to heat as well as terms related to protein folding and chaperone activity (*SI Appendix*, Table S1). Genes down-regulated under stress were enriched in gene ontology terms related to photosynthesis and central metabolism, as expected for mild stress responses.

Despite the large overall changes in gene expression and typical stress-related gene ontology profile, we found surprisingly little intraspecific overlap of gene expression under water-deficit stress in sorghum (Fig. 2 *B* and *C*). Only a single sorghum gene was up-regulated, and no genes down-regulated in all 25 genotypes on the drought sampling date compared with recovery. We defined a set of 269 "shared" differentially expressed genes based on common differential expression between the drought and recovery timepoints in at least half of the sorghum genotypes. Only 133 genes, representing 8% of all up-regulated genes, showed shared upregulation. Similarly, only 136 genes or 6% of down-regulated genes were shared in half or more genotypes. On a per genotype basis, a greater percentage of differential expressed genes were shared, with between 18% and 51% of genes differentially expressed conserved across genotypes. The mean percentage of up-regulated genes in each genotype that were shared across at least half the genotypes (36%) was significantly higher (*t* test $P = 0.01$) than the percentage of shared down-regulated genes (28%). To further explore the difference between shared and variably expressed genes, we defined a set of 1,583 "unique genes" which were differentially expressed in only a single genotype. While the absolute number of unique genes is higher than the shared genes overall, in any given genotype, they represent a lower percentage of the up- and down-regulated genes. We found that the log2 fold-change for shared up-regulated genes was significantly higher compared with unique up-regulated genes (Fig. 2*D*; *t* test $P = 1.48e{-}15$). We compared gene ontology enrichment between the set of shared up-regulated and unique up-regulated genes to ascertain possible differences in function between the two sets. Gene ontology enrichment of the shared up-regulated genes mirrored that of all up-regulated genes, with terms related to response to heat, protein folding, and reactive

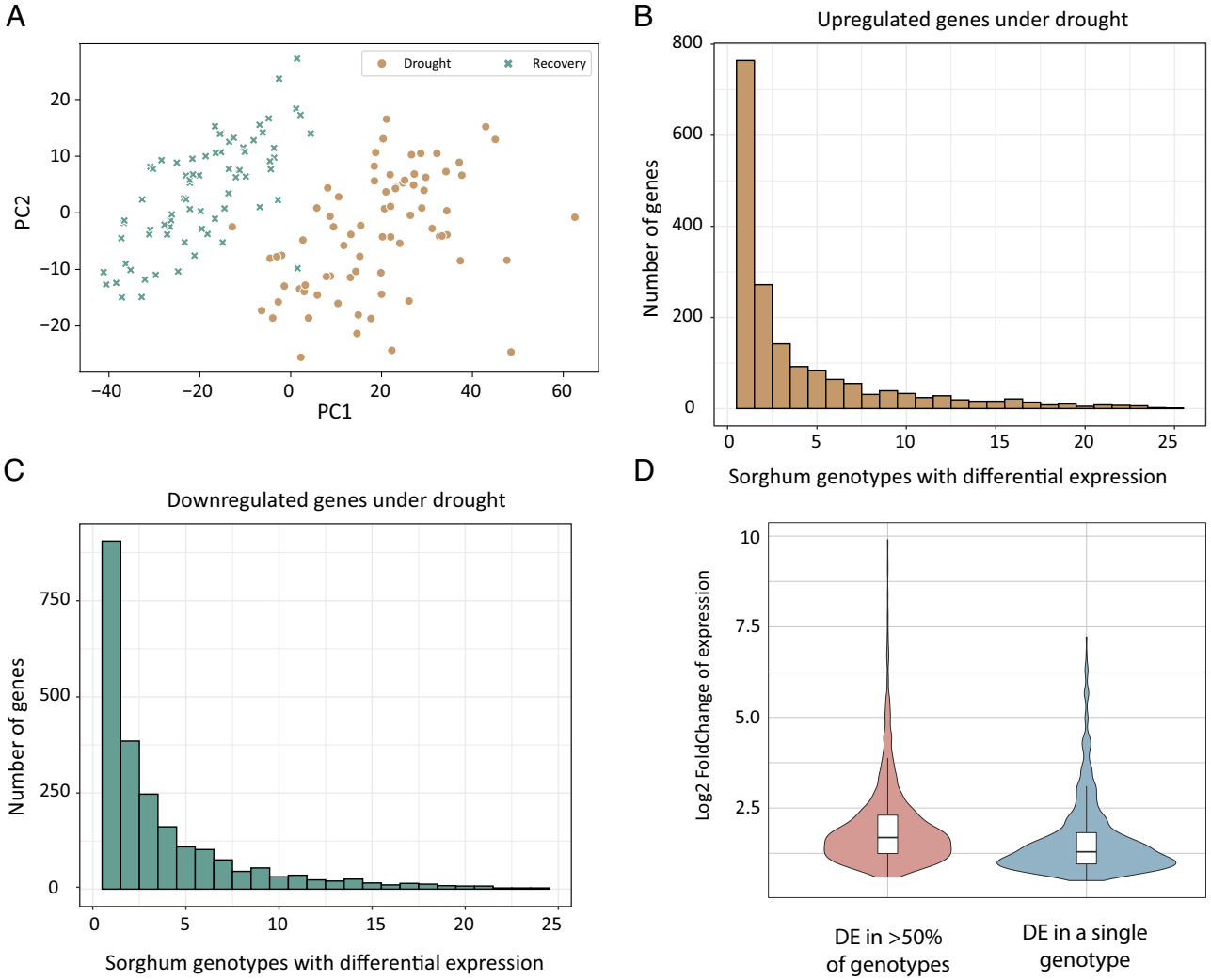

Fig. 2. Unique expression signatures of drought stress across sorghum genotypes. (*A*) Principal component analysis of log2 transformed RNAseq data for the sorghum field drought experiment. Individual samples are plotted and colored by day. Samples are colored by genotype in *SI Appendix*, Fig. S1. (*B*) Histogram showing the number of shared up-regulated expressed genes across the 25 sorghum accessions. (*C*) Histogram showing the number of shared down-regulated expressed genes across the 25 sorghum accessions. (*D*) Violin plots of log2 fold change of expression in the shared differentially expressed genes compared with the genes uniquely differentially expressed in a single genotype.

oxygen species scavenging enriched. Conversely, gene ontology terms enriched among unique up-regulated genes were less obviously related to stress response with terms such as translation, peptide metabolic process, and cellular amide metabolic process enriched.

Predictive Modeling of Drought-Responsive Genes in Sorghum. Our differential expression analysis was limited by either the relatively mild nature of the natural drought event and/or the low number of replicates per genotype (three), potentially causing us to miss shared drought-responsive genes due to insufficient statistical power or a muted drought response due to the mild nature of the stress treatment. We used a predictive modeling approach to more robustly identify any shared water-deficit stress responses across sorghum genotypes in our experiment. Our approach involved training a random forest model to classify samples as "drought" or "control" based on normalized gene expression values alone. We hypothesized that the features with the most predictive power in the model would represent genes with central and conserved roles in drought responses. We first applied this approach to the sorghum experiment described above. Using a training set of expression from 75% of sorghum genotypes, our model had near perfect prediction accuracy on the remaining test set across all folds in a fivefold cross validation scheme. However, the model relied on a small number of features to make those predictions. The average depth of the individual trees within the random forest was only 1.9, indicating that each decision tree used on average less than two genes out of 34,117 possible genes to make a prediction. To improve the utility of our model, we used k-means clustering to reduce the total number of features. We created seven clusters based on the scaled expression data and used the first principal component of gene expression in each cluster as our input feature. Our model was able to classify samples from genotypes withheld from the training data correctly 95% of the time (*SI Appendix*, Fig. S2). The individual trees within the cluster-based model used on average 4.8 out of seven possible k-means-based features. To identify clusters with the most predictive power, we calculated feature importance using the mean decrease in impurity (Gini score) metric implemented in the scikit-learn package. We found that clusters with the most importance in the classification model also had the highest percentage of up-regulated genes (*SI Appendix*, Fig. S3).

Our model was developed using data from a relatively mild drought event, and we expanded the model using publicly available sorghum drought datasets with varying designs, genotypes, and drought severity (*SI Appendix*, Table S2). The public datasets included RNAseq of vegetative tissues from both field and chamber grown sorghum across multiple developmental stages. We also generated an additional dataset using 54 chamber grown Btx623 sorghum plants with drought applied at three different developmental stages. In total, we analyzed seven additional datasets, collectively representing 35 genotypes, with 206 drought-stressed samples and 254 well-watered or recovery samples. We reprocessed all of the expression data using a common analytical framework, and compared these experiments using dimensionality reduction approaches. The expression samples cluster separately by experiment rather than stress vs control along the first two principal components, suggesting significant heterogeneity and sampling artifacts (Fig. 3*A*). To remove batch effects, we applied the combat algorithm to adjust the input data (*SI Appendix*, Fig. S4). We then split the data into training and testing sets using a "leave one experiment out" approach where one dataset was withheld for testing the model and the rest were used for training. The accuracy of our model across all test datasets was 86% (Fig. 3*B*). The precision and recall of the model were both 0.84, where 1 is a perfect classifier and 0.5 is a random classifier (Fig. 3 *C* and *D*). The model performed well across all datasets individually, with prediction accuracy ranging from 64% to 100%; however, excluding the best performing (which only had two samples) and worst performing datasets, the range was 82 to 91% (Fig. 3*B*). The top features (genes) in the sorghum predictive model include 39 of the conserved up-regulated genes from field-stressed plants. The ability of our model to classify samples accurately on an unobserved dataset implies the existence of a conserved pattern of gene expression in response to drought across diverse sorghum lines.

Developmental stage influences the relative drought tolerance of sorghum lines. To assess whether our model could accurately classify drought samples regardless of developmental stage, we trained a version of our model using 203 sorghum samples from the vegetative stage and used the remaining 34 samples from flowering or postflowering stages to test the model. Our model predicted with 97% accuracy and an area under the ROC curve (auc) score of 0.99 (*SI Appendix*, Fig. S5), suggesting that a conserved drought response is present across developmental stages, despite distinct physiological signatures and molecular mechanisms underlying preflowering and postflowering drought responses in sorghum.

Cross-Species Predictive Modeling Identifies Conserved Core Stress Response. Our analyses to this point identified a shared drought response across diverse sorghum lines. To probe the evolutionary conservation of this response, we compared our findings within sorghum to similar datasets in maize. We collected a water-deficit stress dataset across a set of 27 diverse maize genotypes in a greenhouse environment. Briefly, we withheld water from potted maize plants at the ~V5 leaf stage for 1 to 3 d. On each day, we sampled a stressed group and a corresponding control group, which received water daily. We found that stomatal conductance was significantly lower in the stressed groups as compared with the controls ($P = 1.22e{-}105$) as well as across the different experimental timepoints ($P = 1.04e{-}31$) (*SI Appendix*, Fig. S6). As expected, the difference between days was dependent on treatment with a significant interaction between treatment and day ($P = 1.43e{-}38$), demonstrating that the stressed group had a significant drop in stomatal conductance compared with the controls. We also found significant differences in stomatal conductance between genotypes ($P = 0.0017$) suggesting differing physiological responses across maize genotypes.

We observed significant changes in gene expression between the drought and control timepoints in maize, with the number of differentially expressed genes increasing in the more severe timepoint. The maize dataset has only one sample of each genotype at each timepoint, thus instead of comparing differential expression for each genotype, we used log2 fold-change to assess the variation of expression response under water-deficit across genotypes. Similar to the sorghum dataset, we saw limited shared response between the genotypes. Only three genes had a log2 fold-change greater than 1.5 between the first, milder, stress timepoint and the corresponding control in all 27 genotypes. While 272 genes, representing 7% of all up-regulated genes showed a greater than 1.5 fold-change in at least half the genotypes. During the severe stress timepoint, we saw more overlap between genotypes with 125 genes showing a greater than 1.5 fold-change across all genotypes and 1,845 up-regulated in at least half of the genotypes.

We used our modeling approach to test whether sorghum and maize have shared, core drought response pathways. We converted maize genes to their corresponding sorghum orthologs using a synteny-based approach to enable comparisons across species. Although both sorghum and maize share the same

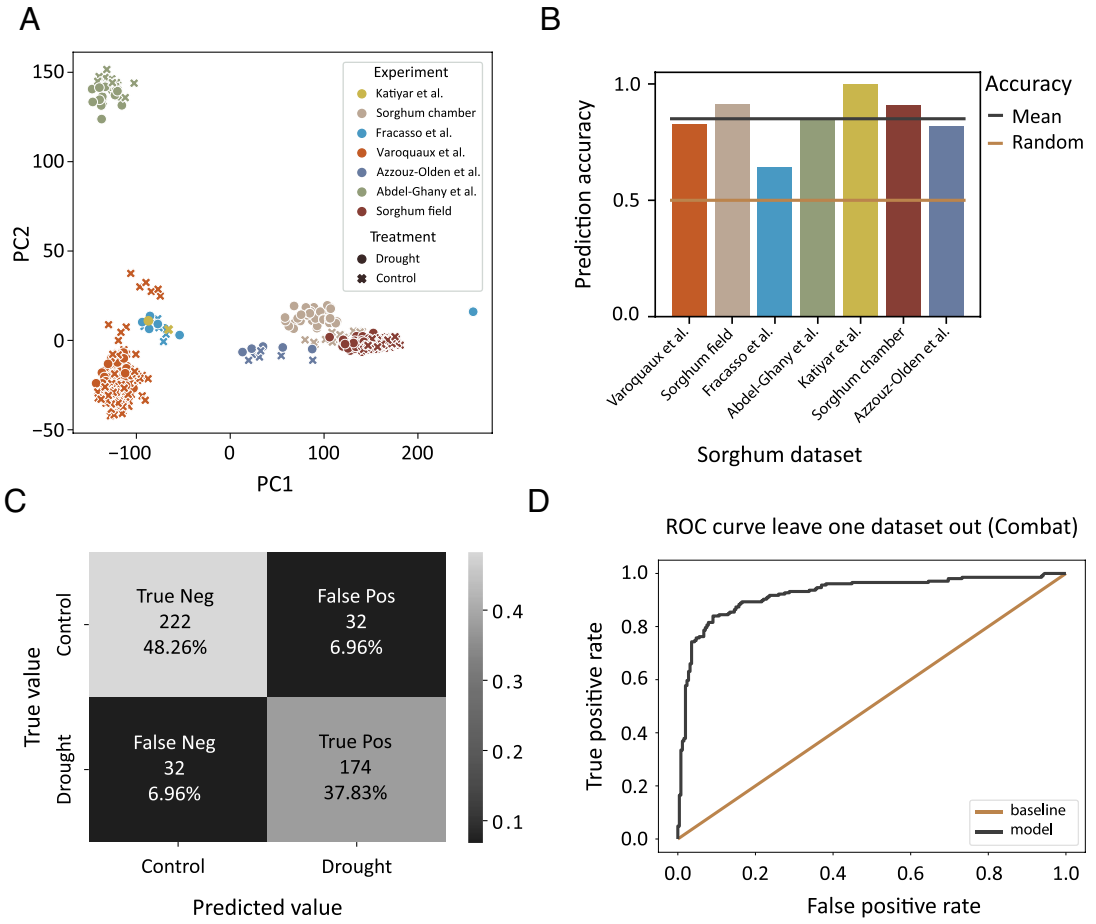

Fig. 3. Predictive modeling of drought stress in sorghum using gene expression data. (*A*) Principal component analysis of log2 transformed RNAseq data for the seven sorghum drought expression datasets used for predictive modeling. A principal component analysis of the ComBat-filtered expression data is available in *SI Appendix*, Fig. S3. (*B*) Predictive accuracy of the random forest model for classifying drought-stressed sorghum samples across each individual experiment. The mean predictive accuracy is shown by a black line compared with a random background (in orange). (*C*) Confusion matrix of the drought predictive model. (*D*) Receiver operating characteristic curve showing the performance of the drought classification model across all classification thresholds.

chromosome number, maize underwent a more recent whole genome duplication and many maize genes display a 2:1 syntenic pattern with sorghum (18). For maize genes with this 2:1 synteny pattern, we averaged syntelog expression and created a converted matrix of maize expression with sorghum gene identifiers. We then retrained our sorghum model using all seven sorghum datasets with only sorghum genes having a syntenic counterpart in maize (Fig. 4*A*). We applied the model to the maize data and found that it predicted with 85% accuracy and an auc score of 0.98. We also created a model trained with the maize data, and tested it on the sorghum dataset. This maize model predicted sorghum samples with 81% accuracy and an auc score of 0.94 across all samples with performances of 64 to 100% across the individual experiments (Fig. 4*B*).

To further test the hypothesis that maize and sorghum share a core stress response, we compared the overlap between differentially expressed genes in the maize dataset, our sorghum field experiment, and the top-predictors from our sorghum-trained model. We found significant overlap between genes up-regulated in maize and sorghum (Fisher's exact test $P = 3.4e{-}47$), as well as genes down-regulated in the two species (Fisher's exact $P = 6.8e{-}158$; Fig. 4 *C* and *D*). We also found significant overlap between the top predictors from the sorghum-trained model and genes up-regulated in maize (Fisher's exact $P = 1.9e{-}47$) as well as sorghum (Fisher's exact $P = 3.5e{-}111$). Interestingly, we did not identify significant overlap between down-regulated genes and the top predictors in our model (Fig. 4*D*).

We hypothesized that expression of the top predictors from our sorghum-trained model would be associated with physiological markers of drought stress. To test this, we calculated the first principal component of log transformed gene expression (PC1) across the top predictors as a summary value of gene expression (*SI Appendix*, Fig. S7). We then correlated the PC1 values with physiological variables. We found that PC1 was significantly negatively correlated with relative water content (Spearman r = −0.53), suggesting that the top predictors are strongly associated with signatures of drought physiology.

We identified a set of 284 genes that were top predictors in the sorghum and maize-trained models, and also differentially expressed in both datasets. The majority of these genes showed increased expression during drought in our sorghum dataset (Fig. 5*A*). Using gene ontology enrichment analysis, we found significant enrichment for well-characterized abiotic and biotic stress-responsive pathways as well as genes related to protein folding (Fig. 5*C*). We found that these conserved drought-responsive genes were also significantly more likely to have shared differential expression (>50% of genotypes) as opposed to differentially expressed in only one sorghum genotype (Fisher's exact $P = 7.37e{-}29$). Previous researchers have identified sets of shared differentially expressed genes related to drought responses in other species (17). To test for overlap between our conserved drought genes in maize and sorghum and across broader species, we used conserved orthogroups to link gene identities between studies. We identified orthologs for

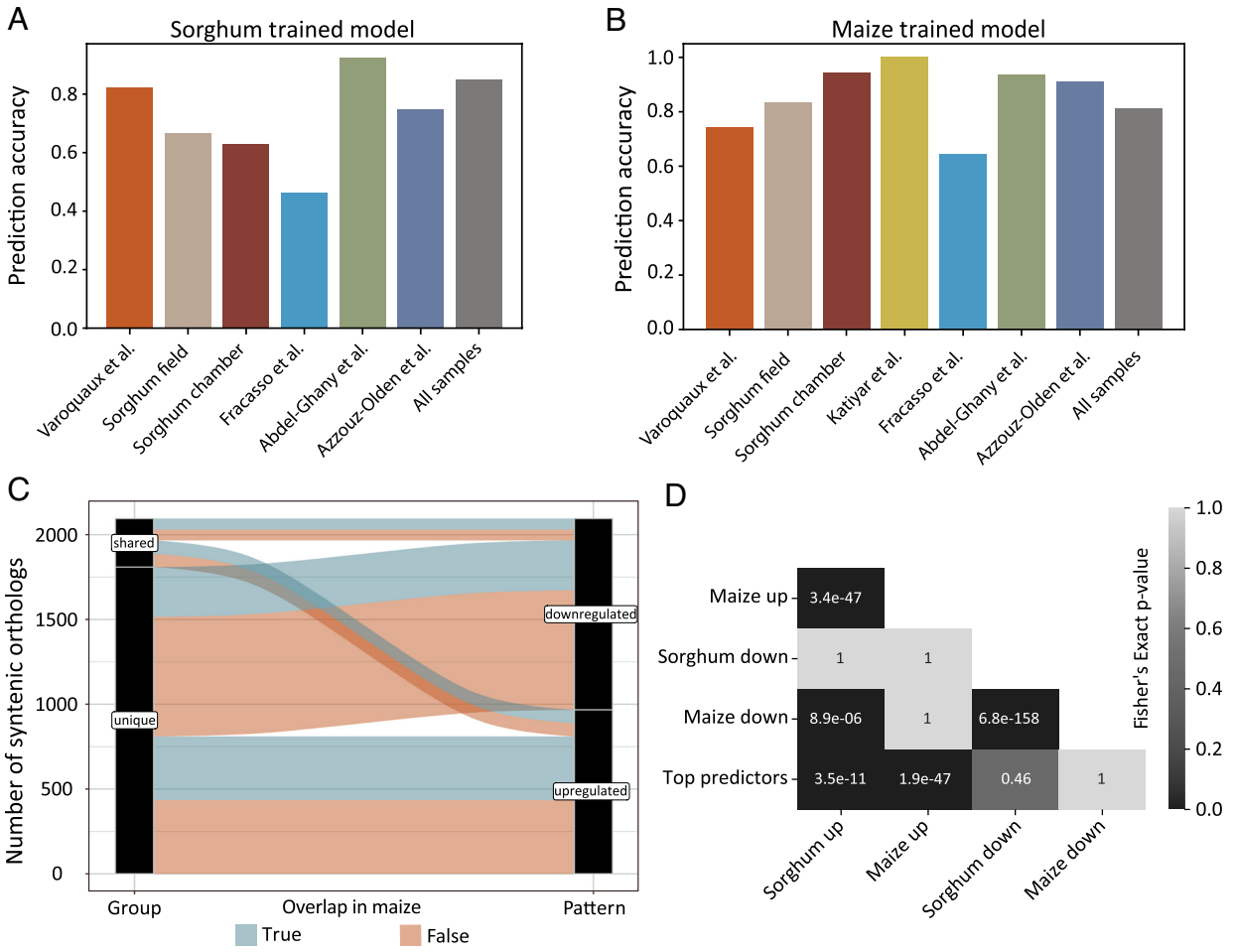

Fig. 4. Cross-species predictive modeling of drought stress. (*A*) Predictive accuracy for classifying drought stress in maize using all of the sorghum samples for training (in gray) and each experiment individually. (*B*) Predictive accuracy of the maize-trained model for classifying drought stress across the sorghum experiments. (*C*) Alluvial plot showing orthologs between maize and sorghum that are conserved top predictors (blue). (*D*) *P*-value from Fisher's exact test comparing overlap between syntenic orthologs in each differentially expressed gene set as well as the top predictors from the sorghum-trained model.

282 of the 284 conserved drought-responsive genes reported in Shaar-Moshe et al. and found 39 had shared drought responsiveness in maize and sorghum. This represents significant enrichment (Fisher's exact test $P = 2.49e{-}20$); however, a substantial portion of the drought response genes between maize and sorghum are not shared with more distantly related species.

Genes with the highest predictive power in the sorghum and maize models are overwhelmingly involved in response to abiotic stresses (*SI Appendix*, Table S3). This include orthologs of the abscisic acid-mediated transcription factors drought-induced protein 19 (Di19-3; Sobic.003G443000) (25) and abscisic acid-responsive element-binding factor 2 (AREB1; Sobic.004G309600) (26), which play central roles in the regulation of drought and high-salinity stress responses. Top predictors also include aquaporin orthologs to plasma membrane intrinsic proteins PIP 1 and 4 (Sobic.004G288700, Sobic.004G238100), which has been linked to various osmotic stresses in Arabidopsis (27). The most abundant group of top predictors are reactive oxygen species scavengers such as L-ascorbate peroxidase (Sobic.001G410200, Sobic.006G084400, Sobic.006G204000), polyamine oxidase (Sobic.001G472000), thioredoxins (Sobic. 001G173500 Sobic.001G386200, Sobic.002G421600, Sobic. 008G117600), ferredoxin 3 (Sobic.001G022900), and oxidoreductase (Sobic.006G140700), as well as various heat shock proteins and a tandem array of orthologs to heat shock protein 22 (HSP22; Sobic.003G081900, Sobic.003G082000, Sobic.003G082100, Sobic.003G082200, Sobic.003G082300, Sobic.003G082500). HSP22s play essential roles in epigenetic memory to heat stress in Arabidopsis (28) and may function in the regulation of osmotic stress. Late embryogenesis-abundant proteins (LEAs) protect cellular macromolecules during water deficit (29), and orthologs of LEA2 (Sobic.001G017100) and dehydrins (Sobic.009G116700) are also included in this list of top predictors. These patterns suggest top predictors are involved in deeply conserved and central responses to water deficit.

Evolutionary Constraint of Conserved Drought-Responsive Genes. Through our predictive modeling approach, we have identified a core set of genes that show a conserved pattern of gene expression during water deficit in maize and sorghum. We expect that the shared expression signatures are an indication of evolutionary conservation. To test this, we compared deleterious load between top predictor genes and a background set of genes. To estimate deleterious load, we used average Sorting Intolerant from Tolerant (SIFT) scores as calculated in Lozano et al. (30). SIFT scores are computational predictions of the effect of individual mutations. A SIFT score below 0.05 represents a mutation that is predicted to be deleterious, and when averaged across all mutations in a gene, the score represents a deleterious index. We compared the proportion of genes with average SIFT scores below 0.05 between 2,000 bootstrapped samples of the

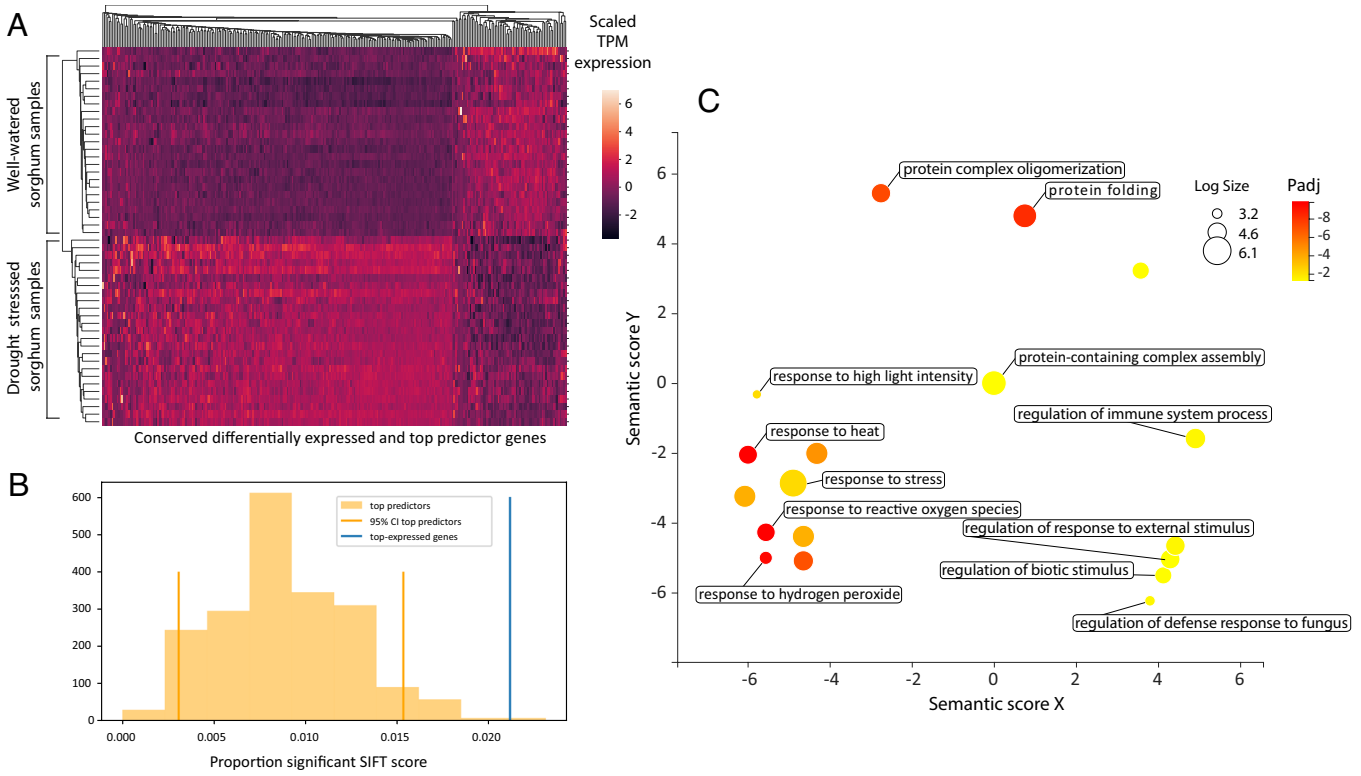

Fig. 5. Evolutionary conservation and functional enrichment of top predictors involved in drought responses. (*A*) Heatmap showing scaled expression values in the sorghum field experiment, for the 284 syntenic orthologs that are differentially expressed in both the maize and sorghum experiments as well as among the top-predictors in the sorghum- and maize-trained models. (*B*) Bootstrapped CI for the proportion of genes with a significant average SIFT score among the sorghum-trained model top predictors compared with the proportion of top expressed genes (>74 percentile of expression) with significant average SIFT scores. (*C*) Multidimensional scaling plot showing clusters of enriched gene ontology terms in the set of genes described above. The size of each circle is proportional to the number of genes annotated with each term, and the circles are colored by the log10 of the adjusted *P*-value.

top predictor genes with a background set. Since genes with high expression are more likely to be evolutionary constrained, we chose the set of all genes with average expression values across all our sorghum field samples greater than the 73rd percentile, which represents the mean percentile rank of the top predictor genes. The top predictor genes had a significantly lower proportion of genes with average SIFT scores < 0.05 than both the highly expressed background set and all genes (Fig. 5*B*). This suggests that the core set of drought-related genes are more evolutionarily constrained than other highly expressed genes.

Discussion

Drought tolerance is variable across diverse sorghum lines, yet some elements of drought response are conserved even across species. Previous work has mostly focused on either differences between individual sorghum genotypes or comparisons of sorghum with other species such as maize. Integrating our understanding of intraspecific and interspecific variations in drought response is an important step in unraveling the evolutionary history of drought tolerance in plants. In this study, we used a predictive modeling approach combined with differential expression analysis across diverse sorghum genotypes to identify shared and unique drought responses.

We identified a core set of genes with a conserved expression pattern across the majority of sorghum genotypes. We then applied our model to a parallel maize dataset and found that the conserved response was largely shared with maize. In evolutionary terms, the ancestors of maize and sorghum diverged relatively recently (18). The two species show conserved response to some stresses, and previous studies have shown conserved resistance mechanisms to particular pathogens between maize and sorghum (31). However, sorghum and maize differ markedly in their resilience to abiotic stresses, particularly drought and heat (19, 32, 33). Interestingly, even for cold stress, an abiotic stress where both species are susceptible, maize and sorghum have surprisingly different gene regulatory responses (34). Therefore, our finding that a core response to drought is conserved between maize and sorghum is initially surprising. A metaanalysis of microarray data identified a set of shared differentially expressed genes across multiple angiosperm species during progressive drought stress, although this work did not include sorghum or maize (17). Our findings expand on this result, showing a similar pattern of conservation in sorghum and maize during drought. Core aspects of angiosperm drought response evolved during the adaptation of early plants to a terrestrial environment (35), and some form of water deficit is continually faced by virtually all plant lineages. Conversely, cold tolerance likely evolved repeatedly across angiosperm lineages and relatively recently in grasses (36). The apparent divergent responses to cold in sorghum and maize and seemingly more shared drought response are perhaps an artifact of the evolution of these two traits across different timescales.

Prior work used differential gene expression to identify shared drought response patterns across species, and this approach has identified hundreds of conserved drought-responsive genes (17). Much of this work was limited to microarray data or small sample sizes that can limit the effectiveness of differential gene expression analysis (37). Combining samples from disparate datasets can increase the sample size; however, this is impractical due to differences in methods between experiments. In particular, drought experiments often represent a broad range of soil water contents, developmental stages, and genotypes. Hundreds of interconnected

pathways underlie a typical drought stress response, and these expression dynamics are not easily captured by simple pairwise differential expression.

Supervised classification models offer an alternative approach to traditional differential gene expression analysis. We used a random forest classifier to label samples as drought or control based on gene expression values. The random forest model was able to accept training data from seven diverse datasets, which varied in sample size, growth environment, developmental stage, and method and level of water-stress imposed. Our model performed well across the majority of these datasets indicating a broadly shared core drought response across disparate sorghum drought treatments. Developmental stage has a major impact on drought tolerance in sorghum, and separate preflowering or postflowering drought tolerant accessions have been identified, with little overlap between groups (5, 9). Preflowering and postflowering drought tolerance strategies are characterized by distinct physiological and molecular mechanisms. Postflowering tolerance is associated with the stay green phenotype where tolerant lines retain green leaf area from anthesis through grain filling. The physiological basis of preflowering drought tolerance is more complex, and likely relates to water use efficiency, osmotic adjustment, and plant architecture traits that ultimately give rise to higher yields (38). Despite these differences, when trained on only the vegetative stage samples, our model still classified flowering and postflowering drought samples accurately. This implies a shared core stress response across developmental stages.

Despite broad conservation of a core set of drought-responsive genes across sorghum datasets and developmental stages, the individual expression response to drought was variable across sorghum genotypes. The majority of differentially expressed genes identified in our sorghum field experiment were private to one genotype. However, within a single genotype, a higher percentage of upregulated genes are shared (36%) than genes which are unique to that genotype (8%). The private genes are potentially responsible for between genotype differences in drought response. Alternatively, they may represent noise or gene expression changes unrelated to drought. Overall, the log2 fold-change of shared genes was significantly higher than unique genes, suggesting that unique genes are more likely to represent noise rather than true differential expression. Furthermore, gene ontology terms enriched among unique genes were not clearly stress related while the shared genes were enriched in terms related to known stress response pathways. While some unique differentially expressed genes are undoubtedly important in drought response, we hypothesize that the core drought response is conserved across genotypes.

Our finding of a conserved drought response across diverse sorghum genotypes and developmental stages coupled with the cross-species predictive accuracy of the sorghum- and maize-trained models suggests an evolutionarily conserved response. A prior metaanalysis found that differentially expressed orthologs which were shared between wheat and rice or barley and rice had higher sequence similarity than orthologs, which were differentially expressed in only one species (17). Not all sequence changes are functionally meaningful. Top predictors in our sorghum-trained model had a significantly lower proportion of genes with average SIFT scores below 0.05 [i.e., predictive of deleterious mutations (39)] than either a random set of background genes or other highly expressed genes. This suggests that conserved drought-responsive genes across sorghum and maize are less likely to contain deleterious mutations.

Several metabolic processes have repeatedly been shown to be involved in drought response across divergent plant species. Gene ontology terms related to response to abiotic stimulus and carbohydrate metabolism were identified as enriched among conserved differentially expressed genes in Shaar-Moshe et al. Other studies proposed that pathways involved in accumulation of osmoprotectants, reactive oxygen species scavenging, regulation of nitrogen metabolism, ammonia detoxification, and activation of the Gamma-aminobutyric acid (GABA) shunt in the tricarboxylic acid (TCA) cycle were conserved across multiple species in response to drought (16). The core sorghum- and maize-responsive genes we identified have overlap between orthogroups identified in Shaar-Moshe et al. and the general gene ontology term "response to abiotic stimulus." We also found evidence of reactive oxygen species scavenging enzymes as well as folding and refolding of proteins based on the gene ontology term enrichment. Cellular response to endoplasmic reticulum stress caused by accumulation of unfolded and misfolded proteins, known as the unfolded protein response (UPR) is a well-studied process in response to environmental stress (40). Much of the UPR is conserved across not just plants but all eukaryotes and thus it is unsurprising that we see shared activation under drought stress here (41).

Conclusion

Prior studies have identified shared differentially expressed genes under drought stress across multiple species. We extend these results and show a similar shared core response across maize and sorghum using a predictive modeling approach. Our approach has the advantage of enabling integration of multiple diverse datasets despite differences in sample size and approach between experiments. We show that the core response is largely shared among diverse sorghum and maize genotypes and across developmental stages despite overall variable drought response between species. Taken together, our results suggest a deeply conserved core drought response exists in plants, and resilience or susceptibility is likely driven by modifications of these central pathways. Practically, the model we created can be used to classify the extent and degree of drought stress experienced by individual plants, and hopefully with additional datasets, could predict signatures that are associated with resilience.

Methods

Sorghum Experimental Design and Sampling. We grew sorghum genotypes from the sorghum association panel for this experiment at the Michigan State University Agronomy farm using a randomized complete block design. The soil type was a mix of Conover loam over approximately two-thirds of the field area, and the more freely draining Sisson fine sandy loam, in the remaining area (US Department of Agriculture, Natural Resources Conservation Service, 2019). We planted seeds in two row plots and allowed the plants to grow under ambient environmental conditions. East Lansing, Michigan, experienced a drier than normal period during the early summer of 2020. The nearby Hancock Turf Research Center weather station recorded only 50.5 mm of precipitation between June 1 and our first sampling date of July 7 compared with the 5-y average of 106.68 mm at that site. The end of June and beginning of July was particularly dry with no precipitation falling between June 27 and the first sampling date of July 7. In total, 42.6 mm of rainfall fell before the second sampling timepoint on July 11.

We sampled each plot at two separate timepoints, the first on July 7, 2020, was during mild water-deficit stress, and the second on July 11 was after the plants had recovered following precipitation. On both days, we took all samples between 10:00 AM and 12:00 PM local time and sky conditions were similar. We collected leaf samples for RNA sequencing into liquid nitrogen from the midsection of the second top-most fully expanded leaf from three plants per plot and combined the samples into a single tube. Leaf tissue from the same leaves were collected into airtight tubes and stored in a cooler for relative water content analysis. We also collected photosynthetic efficiency and other leaf physiology data using the MultiSpeQ fluorometer from the top-most fully expanded leaf for two plants per plot.

We measured the fresh weight (FW) of leaf samples using an analytical balance immediately following field sample collection. We processed three leaf samples from each plot together to achieve a single relative water content value per plot. After measuring fresh weight, we floated the leaf samples in Millipore filtered deionized water kept in the dark overnight at 4 °C. The following day, we dried the surface of the leaf samples and measured the turgid weight (TW) and placed the samples in paper envelopes to dry at 60 °C. After drying overnight, we measured the sample dry weight (DW) and calculated relative water content using the formula:

$$\left(\frac{FW - DW}{TW - DW}\right) * 100\,\%.$$

Maize Experimental Conditions and Sampling. For the maize drought experiment, we grew the 26 founders of the nested association mapping population, as well as the inbred maize line Mo17, in 4″ diameter by 4″ deep nursery pots for three weeks during the month of June 2016 in the Gutterman greenhouse located in Ithaca, NY (42.4482 N, 76.4612 W) (42). Supplemental lighting in the greenhouse provided a minimum of 300 μmol M^{-2} S^{-1} of PAR. During strong sunlight, photosynthetically active radiation (PAR) typically approached 1,000 μmol M^{-2} S^{-1}. The temperature in the greenhouse was held at approximately 28 °C during the day and 20 °C at night. We hand-watered plants twice daily except during drought treatments. We separated the plants into three blocks in the greenhouse with each block consisting of one complete set of genotypes for the control and drought treatments. After 3 wk of growth, at approximately fifth leaf stage, we withheld water from the drought treatment. We measured stomatal conductance on each day of the experiment beginning on day 0 (both the control and drought treatments were well watered) and ending on day 3 (3 d after cessation of water for the drought treatment). A Decagon porometer (model SC-1), calibrated each day of the experiment, was used to collect all stomatal conductance readings from the uppermost fully expanded leaf. All stomatal conductance measurements were collected between 10:00 AM and 2:00 PM local time to minimize the impact of daily physiological cycles on the readings. Pots within the drought treatment were weighed each day of the experiment as a proxy measure for soil moisture. On days 1 and 3 of the experiment, we collected leaf tissue for RNA sequencing in liquid nitrogen. We selected the second top most fully expanded leaf for tissue collection to avoid competition with the leaves selected from physiological measurement. We sampled tissue by folding the leaf from the tip to the base and excising an ∼5-cm section spanning the midpoint of the leaf and extending inward to, but not including the midrib. On day 3, samples were collected from the other half of the second topmost fully expanded leaf when possible. However, in cases where the prior sampling had damaged the leaf, the topmost fully expanded leaf was used as a replacement.

RNAseq Profiling. For both the sorghum and maize experiments, we excised a leaf section from the midpoint of the second top-most fully expanded leaf from three plants per plot (sorghum experiment) or pooled samples from three plants (maize experiment) and froze them in liquid nitrogen. We lysed frozen leaves using a bead tissue homogenizer. We then thawed the ground tissue in TRIzol reagent and extracted RNA using a Direct-zol 96 kit according to the manufacturer's instructions (Zymo Research, Irvine, CA). Lexogen quant-seq libraries for each sample were prepared and sequenced by the Cornell Institute of Biotechnology for the maize and sorghum datasets.

Previously published RNAseq data for drought stress in sorghum were collected from refs. 10–12, 21, and 22 and downloaded from the National Center for Biotechnology Information (NCBI) sequence read archive and processed as described below. Full details of the published RNAseq data can be found in *SI Appendix*, Table S2.

RNA Sequence Processing. We trimmed sequence adapters and quality checked the raw FASTQ files using the program fastp (v0.23.2) (43). We then pseudoaligned our cleaned sequencing reads to the Btx623 sorghum or B73 V5 maize reference genomes using salmon (v1.6.) (44–46). We then converted transcript level counts to gene level using the R package TXimport (v 1.22.0) (47). We used DESeq2 (v1.36.0) to calculate pairwise differential expression between drought and well-watered conditions for each genotype (48).

RNAseq Data Normalization and Batch Effect Removal. For our model built across sorghum datasets, we removed batch effects using the combat algorithm implemented in the python package pyComBat (v0.3.2) (49). For all models, we split the data into training and testing sets using approaches outlined in *SI Appendix*, Table S4. After splitting the data into training and test sets, we scaled the data using the StandardScalar function from the scikit-learn python package (v1.1.0) (50).

Random Forest Model Construction and Feature Importance. We constructed random forest models with the RandomForestClassifier function from scikit-learn (v1.1.0) (50). To select hyper-parameters, we used the RandomizedGridSearchCV function with 100 iterations using threefold cross validation to search the parameter space (*SI Appendix*, Table S5).

We calculated feature importance using mean decrease in impurity (Gini score) as implemented in the scikit-learn package (v1.1.0). We then ranked all genes by their importance score. To identify a set number of "top predictors" we used a heuristic approach whereby we selected the *n* top features and compared the number that overlapped with differentially expressed genes with the number of overlaps in a random set of *n* genes. For each set of size *n*, we calculated a z-score (# of *n* top predictors that are also differentially expressed–mean (# of *n* randomly selected genes that are differentially expressed)/SD of random genes. We then selected 675 top predictor genes, as that maximized the z-score.

Data, Materials, and Software Availability. RNAseq data generated in this project are available on the NCBI sequence read archive for maize and sorghum under BioProject PRJNA906711. All other data are included in the manuscript and/or *SI Appendix*. Previously published RNAseq data for drought stress in sorghum was collected from refs. 10–12, 21, and 22 and downloaded from the NCBI sequence read archive and processed as described below. Full details of the published RNAseq data can be found in *SI Appendix*, Table S2.

ACKNOWLEDGMENTS. This work is supported by NSF MCB–1817347 and the United States Department of Agriculture National Institute of Food and Agriculture (USDA-NIFA 2022-67013-36118) to R.V. M.H. was a participant in the Plant Genomics Research Experience for Undergraduates Program funded by NSF Division of Biological Infrastructure (NSF-DBI 1358474). J.P. was supported by predoctoral training award T32-GM110523 from the National Institute of General Medical Sciences of the NIH.

Author affiliations: [a]Department of Horticulture, Michigan State University, East Lansing, MI 48824; [b]Plant Resilience Institute, Michigan State University, East Lansing, MI 48824; [c]Department of Plant Biology, Michigan State University, East Lansing, MI 48824; [d]Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853; [e]School of Integrative Plant Science, Cornell University, Ithaca, NY 14853; [f]Agricultural Research Service, US Department of Agriculture, Ithaca, NY 14853; and [g]Department of Plant, Soil and Microbial Sciences, Michigan State University, East Lansing, MI 48824

1. M. J. Hayes, M. D. Svoboda, B. D. Wardlow, M. C. Anderson, F. Kogan, "Drought monitoring: Historical and current perspectives" in *Drought Mitigation Center Faculty Publications* (University of Nebraska, Lincoln, NE, 2012).
2. M. Ilyas *et al.*, Drought Tolerance strategies in plants: A mechanistic approach. *J. Plant Growth Regul.* **40**, 926–944 (2021).
3. T. Umezawa, M. Fujita, Y. Fujita, K. Yamaguchi-Shinozaki, K. Shinozaki, Engineering drought tolerance in plants: Discovering and tailoring genes to unlock the future. *Curr. Opin. Biotechnol.* **17**, 113–122 (2006).
4. X. Feng *et al.*, The ecohydrological context of drought and classification of plant responses. *Ecol. Lett.* **21**, 1723–1736 (2018).
5. A. J. Ogden, S. Abdali, K. M. Engbrecht, M. Zhou, P. P. Handakumbura, Distinct preflowering drought tolerance strategies of sorghum bicolor genotype RTx430 revealed by subcellular protein profiling. *Int. J. Mol. Sci.* **21**, 9706 (2020).
6. J. Pardo, R. VanBuren, Evolutionary innovations driving abiotic stress tolerance in C4 grasses and cereals. *Plant Cell* **33**, 3391–3401 (2021).
7. K. Venkateswaran, M. Elangovan, N. Sivaraj, "Chapter 2–Origin, domestication and diffusion of sorghum bicolor" in *Breeding Sorghum for Diverse End Uses*, C. Aruna, K. B. R. S. Visarada, B. V. Bhat, V. A. Tonapi, Eds. (Woodhead Publishing, 2019), pp. 15–31.
8. D. T. Rosenow, J. E. Quisenberry, C. W. Wendt, L. E. Clark, Drought tolerant sorghum and cotton germplasm. *Agric. Water Manage.* **7**, 207–222 (1983).
9. K. Harris *et al.*, Sorghum stay-green QTL individually reduce post-flowering drought-induced leaf senescence. *J. Exp. Bot.* **58**, 327–338 (2007).
10. N. Varoquaux *et al.*, Transcriptomic analysis of field-droughted sorghum from seedling to maturity reveals biotic and metabolic responses. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 27124–27132 (2019), 10.1073/pnas.1907500116.

11. A. Fracasso, L. M. Trindade, S. Amaducci, Drought stress tolerance strategies revealed by RNA-Seq in two sorghum genotypes with contrasting WUE. *BMC Plant Biol.* **16**, 115 (2016).
12. S. E. Abdel-Ghany, F. Ullah, A. Ben-Hur, A. S. N. Reddy, Transcriptome analysis of drought-resistant and drought-sensitive sorghum (Sorghum bicolor) genotypes in response to PEG-induced drought stress. *Int. J. Mol. Sci.* **21**, 772 (2020).
13. S. M. Johnson, I. Cummins, F. L. Lim, A. R. Slabas, M. R. Knight, Transcriptomic analysis comparing stay-green and senescent Sorghum bicolor lines identifies a role for proline biosynthesis in the stay-green trait. *J. Exp. Bot.* **66**, 7061–7073 (2015).
14. J. L. Boatwright *et al.*, Sorghum association panel whole-genome sequencing establishes cornerstone resource for dissecting genomic diversity. *Plant J.* **111**, 888–904 (2022).
15. J. E. Spindel *et al.*, Association mapping by aerial drone reveals 213 genetic associations for Sorghum bicolor biomass traits under drought. *BMC Genom.* **19**, 679 (2018).
16. R. C. Rabara *et al.*, Tobacco drought stress responses reveal new targets for Solanaceae crop improvement. *BMC Genom.* **16**, 484 (2015).
17. L. Shaar-Moshe, S. Hübner, Z. Peleg, Identification of conserved drought-adaptive genes using a cross-species meta-analysis approach. *BMC Plant Biol.* **15**, 111 (2015).
18. Z. Swigoňová *et al.*, Close split of sorghum and maize genome progenitors. *Genome Res.* **14**, 1916–1923 (2004).
19. S. Schittenhelm, S. Schroetter, Comparison of drought tolerance of maize, sweet sorghum and sorghum-sudangrass hybrids. *J. Agron. Crop Sci.* **200**, 46–53 (2014).
20. D. Ortiz, M. G. Salas-Fernandez, Dissecting the genetic control of natural variation in sorghum photosynthetic response to drought stress. *J. Exp. Bot.* **73**, 3251–3267 (2022).
21. F. Azzouz-Olden, A. G. Hunt, R. Dinkins, Transcriptome analysis of drought-tolerant sorghum genotype SC56 in response to water stress reveals an oxidative stress defense strategy. *Mol. Biol. Rep.* **47**, 3291–3303 (2020).
22. A. Katiyar *et al.*, Identification of novel drought-responsive microRNAs and trans-acting siRNAs from Sorghum bicolor (L.) Moench by high-throughput sequencing analysis. *Front. Plant Sci.* **6**, 506 (2015).
23. S. Tietz, C. C. Hall, J. A. Cruz, D. M. Kramer, NPQ(T): A chlorophyll fluorescence parameter for rapid estimation and imaging of non-photochemical quenching of excitons in photosystem-II-associated antenna complexes. *Plant Cell Environ.* **40**, 1243–1255 (2017).
24. J. Zhuang *et al.*, Drought stress strengthens the link between chlorophyll fluorescence parameters and photosynthetic traits. *PeerJ* **8**, e10046 (2020).
25. L.-X. Qin *et al.*, Arabidopsis drought-induced protein Di19-3 participates in plant response to drought and high salinity stresses. *Plant Mol. Biol.* **86**, 609–625 (2014).
26. T. Yoshida *et al.*, Four*Arabidopsis* AREB/ABFtranscription factors function predominantly in gene expression downstream ofSnRK2kinases in abscisic acid signalling in response to osmotic stress. *Plant Cell Environ.* **38**, 35–49 (2015).
27. J. Y. Jang *et al.*, Transgenic Arabidopsis and tobacco plants overexpressing an aquaporin respond differently to various abiotic stresses. *Plant Mol. Biol.* **64**, 621–632 (2007).
28. N. Yamaguchi *et al.*, H3K27me3 demethylases alter HSP22 and HSP17.6C expression in response to recurring heat in Arabidopsis. *Nat. Commun.* **12**, 3480 (2021).
29. S. C. Hand, M. A. Menze, M. Toner, L. Boswell, D. Moore, LEA proteins during water stress: Not just for plants anymore. *Annu. Rev. Physiol.* **73**, 115–134 (2011).
30. R. Lozano *et al.*, Comparative evolutionary genetics of deleterious load in sorghum and maize. *Nat. Plants* **7**, 17–24 (2021).
31. X. Zhang *et al.*, Conserved defense responses between maize and sorghum to Exserohilum turcicum. *BMC Plant Biol.* **20**, 67 (2020).
32. L. Busta, E. Schmitz, D. K. Kosma, J. C. Schnable, E. B. Cahoon, A co-opted steroid synthesis gene, maintained in sorghum but not maize, is associated with a divergence in leaf wax chemistry. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2022982118 (2021).
33. S. Choudhary *et al.*, Maize, sorghum, and pearl millet have highly contrasting species strategies to adapt to water stress and climate change-like conditions. *Plant Sci.* **295**, 110297 (2020).
34. Y. Zhang *et al.*, Differentially regulated orthologs in sorghum and the subgenomes of maize. *Plant Cell* **29**, 1938–1951 (2017).
35. C. Zhao *et al.*, Evolution of chloroplast retrograde signaling facilitates green plant adaptation to land. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 5015–5020 (2019).
36. A. M. Humphreys, H. P. Linder, Evidence for recent evolution of cold tolerance in grasses suggests current distribution is not limited by (low) temperature. *New Phytol.* **198**, 1261–1273 (2013).
37. C. Stretch *et al.*, Effects of sample size on differential gene expression, rank order and prediction accuracy of a gene signature. *PLoS One* **8**, e65380 (2013).
38. M. R. Tuinstra, E. M. Grote, P. B. Goldsbrough, G. Ejeta, Identification of quantitative trait loci associated with pre-flowering drought tolerance in sorghum. *Crop Sci.* **36**, 1337–1344 (1996).
39. P. C. Ng, S. Henikoff, SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
40. Y.-S. Lai *et al.*, Systemic signaling contributes to the unfolded protein response of the plant endoplasmic reticulum. *Nat. Commun.* **9**, 3918 (2018).
41. L. Zhang, C. Zhang, A. Wang, Divergence and conservation of the major UPR branch IRE1-bZIP signaling pathway across eukaryotes. *Sci. Rep.* **6**, 27362 (2016).
42. M. D. McMullen *et al.*, Genetic properties of the maize nested association mapping population. *Science* **325**, 737–740 (2009).
43. S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
44. M. B. Hufford *et al.*, De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science* **373**, 655–662 (2021).
45. R. F. McCormick *et al.*, The Sorghum bicolor reference genome: Improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *Plant J.* **93**, 338–354 (2018).
46. R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, C. Kingsford, Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
47. C. Soneson, M. I. Love, M. D. Robinson, Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences. *F1000Res.* **4**, 1521 (2015).
48. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
49. A. Behdenna, J. Haziza, C.-A. Azencott, A. Nordor, pyComBat, a Python tool for batch effects correction in high-throughput molecular data using empirical Bayes methods. bioRxiv [Preprint] (2021). https://doi.org/10.1101/2020.03.17.995431 (Accessed 15 February 2023).
50. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, Scikit-learn: Machine learning in Python. *J. Mach. Learn.* **12**, 2825–2830 (2011).