Extensive modulation of a conserved *cis*-regulatory code across 589 grass species

AUTHORS: Charles O. Hale[1*], Sheng-Kai Hsu[2], Jingjing Zhai[2], Aimee J. Schulz[1], Taylor Aubuchon-Elder[4], Germano Costa-Neto[2], Allen Gelfond[1], Mohamed El-Walid[1], Matthew Hufford[5], Elizabeth A. Kellogg[4], Thuy La[2], Alexandre P. Marand[6], Arun S. Seetharam[5], Armin Scheben[7], Michelle Stitzer[2], Travis Wrightsman[1], M. Cinta Romay[2], Edward S. Buckler[1,2,3]

[1]Section of Plant Breeding and Genetics, Cornell University, Ithaca, NY USA 14853

[2]Institute for Genomic Diversity, Cornell University, Ithaca, NY USA 14853

[3]USDA-ARS; Ithaca, NY, USA 14853

[4]Donald Danforth Plant Science Center, St. Louis, MO USA 63132

[5]Iowa State University, Ames, IA USA 50011

[6]University of Michigan, Ann Arbor, MI USA 48109

[7]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY USA 11724

*Correspondence: Charles Hale (coh22@cornell.edu)

ABSTRACT

The growing availability of genomes from non-model organisms offers new opportunities to identify functional loci underlying trait variation through comparative genomics. While *cis*-regulatory regions drive much of phenotypic evolution, linking them to specific functions remains challenging. We identified 514 *cis*-regulatory motifs enriched in regulatory regions of five diverse grass species, with 73% consistently enriched across all, suggesting a deeply conserved regulatory code. We then quantified conservation of specific motif instances across

589 grass species, revealing widespread gain and loss over evolutionary time. Conservation declined rapidly over the first few million years of divergence, yet ~50% of motif instances were conserved back to the origin of grasses ~100 million years ago. Conservation patterns varied by gene class, with modestly higher conservation at transcription factor genes. To test for adaptive *cis*-regulatory changes, we used phylogenetic mixed models to identify motif gains and losses associated with ecological niche transitions. Our models revealed polygenic adaptation across 810 motif-orthogroup combinations, including convergent gains of HSF/GARP motifs at an Alpha-N-acetylglucosaminidase gene associated with adaptation to temperate environments. Our results support a model in which *cis*-regulatory evolution involves extensive turnover of individual binding site instances while largely preserving transcription factors' binding preferences. Convergent *cis*-regulatory changes at hundreds to thousands of genes likely contribute to environmental adaptation. Our results highlight the potential of comparative genomics and phylogenetic mixed models to reveal the genetic basis of complex traits.

INTRODUCTION

Cis-regulatory changes are arguably the most important genetic mechanism of evolutionary innovation, particularly over relatively shallow evolutionary time scales (King and Wilson 1975; Carroll 2008). While the evolution of *trans*-acting regulators such as transcription factors (TFs) typically have widespread effects genome-wide (Signor and Nuzhdin 2018), the evolution of *cis*-regulatory regions enables finely-tuned gene expression evolution at individual genes (Wittkopp et al. 2004; Prud'homme et al. 2007; Wray 2007). Variation within noncoding regions explains roughly half of additive trait variance within maize populations (Rodgers-Melnick et al.

2016), underscoring the importance of these regions in driving phenotypic evolution within species.

TF binding sites (TFBS) are key components of non-coding regulatory regions (Schmitz et al. 2021). TFBS are typically composed of a core DNA sequence motif that is recognized and bound by its cognate transcription factor. In plants, TFBS variants have been shown to be associated with the evolution of traits including inflorescence branching (Hendelman et al. 2021), abiotic stress tolerance (Jiang et al. 2022; Zeng et al. 2025) and fruit shape (Hu et al. 2024). Variants in the 5' UTR and proximal promoter region are often particularly important for modulating expression levels (Cui et al. 2023; Voichek et al. 2024). Despite the importance of these TF binding variants in driving phenotypic evolution, characterizing key variants remains a major challenge. TF binding assays such as Chromatin Immunoprecipitation Sequencing (ChIP-seq), MNase-defined cistrome-Occupancy Analysis (MOA-seq), and DNA affinity purification sequencing (DAP-seq) have been very helpful for precisely characterizing where TFs bind throughout plant genomes (O'Malley et al. 2016; Savadel et al. 2021; Marand et al. 2022), but adapting them to work in non-model systems can be difficult. In particular, the increasing number of sequenced taxa requires scalable methods to discover key *cis*-regulatory features.

Characterizing evolutionary constraint and convergence across large numbers of taxa has emerged as a powerful strategy for determining genetic function from DNA sequence alone (Smith et al. 2020). The identification of conserved noncoding sequences has been used to identify highly important *cis*-regulatory features across non-model taxa (Haudry et al. 2013; Song et al. 2021; Stitzer et al. 2025). A limitation of this approach, however, is that many key *cis*-regulatory regions cannot be aligned reliably, even when function is shared (Schmidt et al.

2010; Yang et al. 2015; Kaplow et al. 2023). Alignment-free characterization approaches can offer a broader view of *cis*-regulatory evolution (Kaplow et al. 2023).

Due to these advances, our understanding of *cis*-regulatory evolution and its phenotypic implications is rapidly increasing. Still, some of the fundamental principles of regulatory evolution remain incompletely understood. A growing body of evidence indicates that TFs and their binding site preferences remain largely intact within lineages, suggesting a conserved "regulatory code" (Stergachis et al. 2014; Nitta et al. 2015; Tu et al. 2020). Changes to the regulatory code can occur via expansions or contractions of TF families that lead to the birth or extinction of particular TF binding preferences (Lehti-Shiu et al. 2017). Across deep evolutionary divergence, such as between plants and animals, widespread differences are observed between TFs and their binding preferences (Riechmann et al. 2000). The field lacks a clear consensus on the extent to which changes to TF binding preferences occur over varying time scales, as well as how and when changes occur. A deeper understanding of regulatory code evolution will inform how to effectively transfer genetics across species.

Another emerging consensus is that while TF binding preferences tend to be deeply conserved, individual instances of TF binding sites are extensively gained and lost over evolutionary time. Previous studies have estimated that only 20% of TFBS instances are conserved between humans and mice (Stergachis et al. 2014), and 20-40% of GOLDEN2-LIKE binding sites were found to be conserved between maize and rice (Tu et al. 2022). Relatively few studies so far have traced TFBS evolution across a large number of taxa (but see (Andrews et al. 2023)), or associated particular gain and loss events with diversification and adaptation at the macroevolutionary scale.

Grasses (Poaceae) have diversified widely over roughly 100 million years of evolution (Gallaher et al. 2022), occupying habitats ranging from tropical savannas to the Arctic tundra (Bouchenak-Khelladi et al. 2010; Edwards and Smith 2010; Spriggs et al. 2014; Lehmann et al.). Grasses' ecological diversification has been intertwined with repeated innovations in life history strategies (annual vs. perennial) (Kellogg 2015), photosynthetic pathway transitions (Grass Phylogeny Working Group II 2012), and photoperiod changes (Fjellheim et al. 2014). Despite these significant shifts, aspects of genome structure such as gene content and collinearity are largely preserved across grass lineages (Bennetzen and Freeling 1993; McSteen and Kellogg 2022; Mascher et al. 2024), although ploidy (Stebbins 1985; Zhang et al. 2024), genome size (Bennett and Smith 1976) and gene regulatory patterns (Meng et al. 2021; Stitzer et al. 2025) vary widely across the family. Comparative genomic analyses across grasses are therefore well-suited to illuminate how broad and repeated environmental adaptation unfolds in complex genomes (Buell 2009).

We hypothesized that extensive and repeated turnover of individual transcription factor binding site (TFBS) instances—rather than changes in transcription factor binding preferences—underlies grass diversification, enabling grasses to fine-tune gene regulation while maintaining regulatory logic. To test these hypotheses, we performed large-scale comparative genomic analyses across a collection of 727 genome assemblies representing 589 grass species. After establishing that grasses share a conserved set of TF motifs enriched in *cis*-regulatory regions, we quantified gain and loss of motif instances across species and performed association mapping across species to identify examples of motif gain and loss associated with environmental adaptation. We documented widespread gain and loss of motif instances as grasses adapted and diversified. Additionally, we revealed 810 environmentally associated motif

gain/loss instances with moderate to weak levels of convergence. Together, our findings support a model of *cis*-regulatory evolution in which adaptation proceeds over macroevolutionary time scales via hundreds to thousands of *cis*-regulatory changes at small-effect downstream genes rather than at a few large-effect regulators.

RESULTS

Grasses share a deeply conserved *cis*-regulatory code

We hypothesized that a highly similar repertoire of *cis*-regulatory motifs are enriched in *cis*-regulatory regions across grasses. We estimated *cis*-regulatory regions using unmethylated regions (UMRs) available from five species (*Brachypodium distachyon, Oryza sativa*, *Sorghum bicolor*, *Zea mays*, and *Hordeum vulgare*). These five species last shared a common ancestor before the BOP (Bambusoideae, Oryzoideae, and Pooideae) and PACMAD (Panicoideae, Aristidoideae, Chloridoideae, Micrairoides, Arundinoideae, and Danthonioideae) clades diverged approximately 80 million years ago (Gallaher et al. 2022) (Fig. 1a). UMRs stably mark functional regulatory and genic regions of plant genomes and are rich in transcription factor binding sites (Crisp et al. 2020). To determine which *cis*-regulatory motifs are commonly enriched in regulatory regions, we quantified UMR enrichment of 704 experimentally-derived plant TF binding motifs from the JASPAR 2024 database (Rauluseviciute et al. 2023). Using randomized UMR sequences with preserved dinucleotide frequencies as background, we measured enrichment of each JASPAR motif in true UMRs relative to dinucleotide-shuffled UMRs.

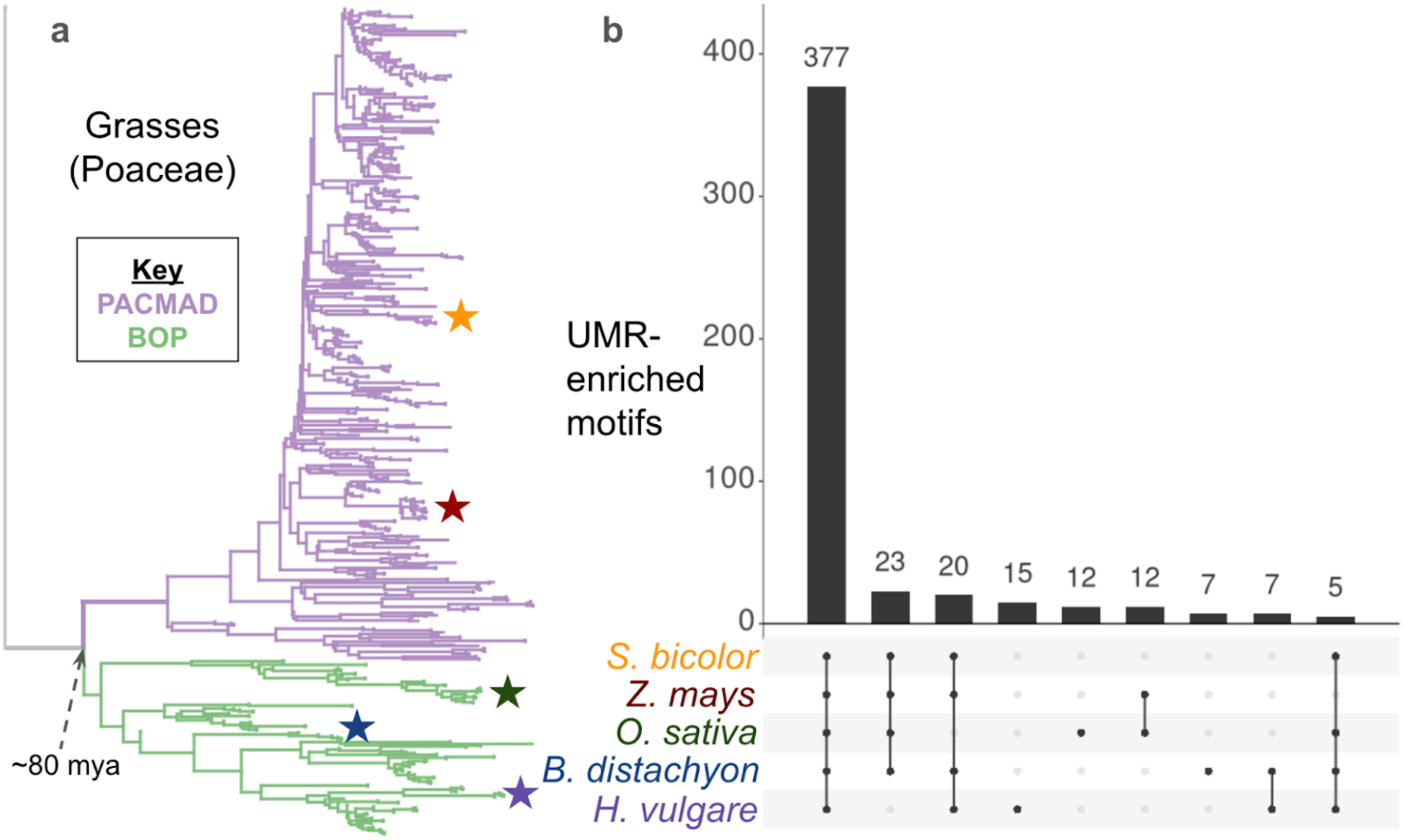

Figure 1. (a) Phylogenetic tree of 589 Poaceae species. Five representative species are starred: *Sorghum bicolor* (orange), *Zea mays* (red), *Oryza sativa* (green), *Brachypodium distachyon* (blue), *Hordeum vulgare* (purple). (b) Enrichment of transcription factor motifs in unmethylated regions across five representative grass species. The intersection bars show the number of enriched motifs for each species set. Intersections with fewer than five motifs are not shown.

Enrichment fold-change showed correlations of 96-99% between *B. distachyon, O. sativa,* and *S. bicolor*, as well as with *Zea mays* and *Hordeum vulgare* (Fig. S1). Of the 514 motifs that were enriched in at least one species, 73% (377 / 514) were commonly enriched across all five species (Fig. 1), with a similar set of motifs enriched in accessible chromatin regions across species (Fig. S2). We used this set of 377 commonly enriched motifs (Supplementary Table 1) for subsequent comparisons of motif conservation and turnover across the grass family.

Patterns of motif occurrence across 589 widely-adapted species

To characterize *cis*-regulatory evolution on a large scale, we built a pipeline to quantify and compare motif occurrences across orthologous regions of hundreds of species (Fig. 2a). For this study, we used publicly-available WGS short reads to generate genome assemblies for 57 species. While not highly contiguous, our WGS assemblies recovered a median of 4206 genes (75%) from a set of 5592 conserved grass genes in a BUSCO-like analysis (Fig. S3b, Supplementary Table 2). In total, we amassed a dataset of 727 genome assemblies representing 589 diverse grass species using 217 public genome assemblies, 427 newly generated short read assemblies from Schulz et al (unpublished data), 58 short read assemblies from (Schulz et al. 2023) and 33 highly contiguous assemblies from (Stitzer et al. 2025) (Fig. S3a, Supplementary Table 2). Our dataset captures wide and repeated environmental adaptation across the grass family (Hsu et al, unpublished data).

To enable cross-species comparisons, we identified orthologous genes across taxa by querying a set of ancestrally reconstructed protein sequences against each assembly and retaining the primary alignment. We retained 15,555 orthogroups after filtering, each containing an average of 465 taxa (SD = 104) (Fig. S4), indicating a high level of taxonomic representation suitable for robust comparative analyses. We scanned intervals 500 bp upstream of the aligned translation start sites for the set of 377 motifs with conserved UMR enrichment. While many important regulatory elements are present further upstream (or downstream) from genes, we considered only 500 bp upstream due to 1) limited contig length in our short-read assemblies and 2) the higher proportion of transposable elements present further upstream. We collapsed overlapping instances of similar motifs into a single merged interval to reduce redundancy for subsequent analyses, using 35 motif clusters delineated by JASPAR2024 (Supplementary Table

171 1).

172 The 35 motif cluster types differed widely in abundance and variability across taxa (Fig
173 2b). Most motif clusters had a median occurrence rate of ~0.25 instances per upstream region,
174 but some motif clusters such as those containing ethylene-responsive factors (ERF) and
175 dehydration responsive element binding proteins (DREB) were highly abundant, appearing
176 upwards of twice per upstream region on average. Occurrence rate variability across species also
177 differed widely across motif clusters, with coefficients of variation ranging from 0.03 to 0.14
178 (Fig 2b).

179 One challenge of *de novo* TFBS prediction using motifs is that many false positives are
180 typically observed when scanning genome-wide (Grant et al. 2011; Jayaram et al. 2016).
181 Restricting motif scanning to regions with external functional evidence (e.g. regions with
182 evolutionary conservation, open chromatin, or in this case, gene proximity) can reduce the
183 number of false positives (Wasserman and Sandelin 2004), though motifs remain unreliable
184 predictors of *in vivo* binding (Mahendrawada et al. 2025). To investigate overlap of our set of
185 gene-proximal motifs with known regulatory regions, we quantified the proportion of our maize
186 motif instances found within independently generated maize MOA-seq (TF binding) peaks
187 (Engelhorn et al. 2023), accessible chromatin regions from a single-cell assay for transposase
188 accessible chromatin with sequencing (scATAC-seq) (Marand et al. 2021), and conserved
189 noncoding sequences (CNS) (Stitzer et al. 2025). The majority of maize motifs intersected
190 scATAC-seq (62.1%) and MOA-seq peaks (69.8%), while only 26.8% intersected CNS (Fig. 2c).
191 While our focus on proximal upstreams prevented us from examining the multitude of distal TF
192 binding sites that can be captured by CNS and ‘omics approaches, our approach captures a set of
193 motifs in active regulatory regions that are not alignable and may be missed by CNS analyses.

For example, we were able to recover a known functional MYB TFBS at *ZmICE1* (Jiang et al. 2022) that is not found within a CNS (Fig. S5).

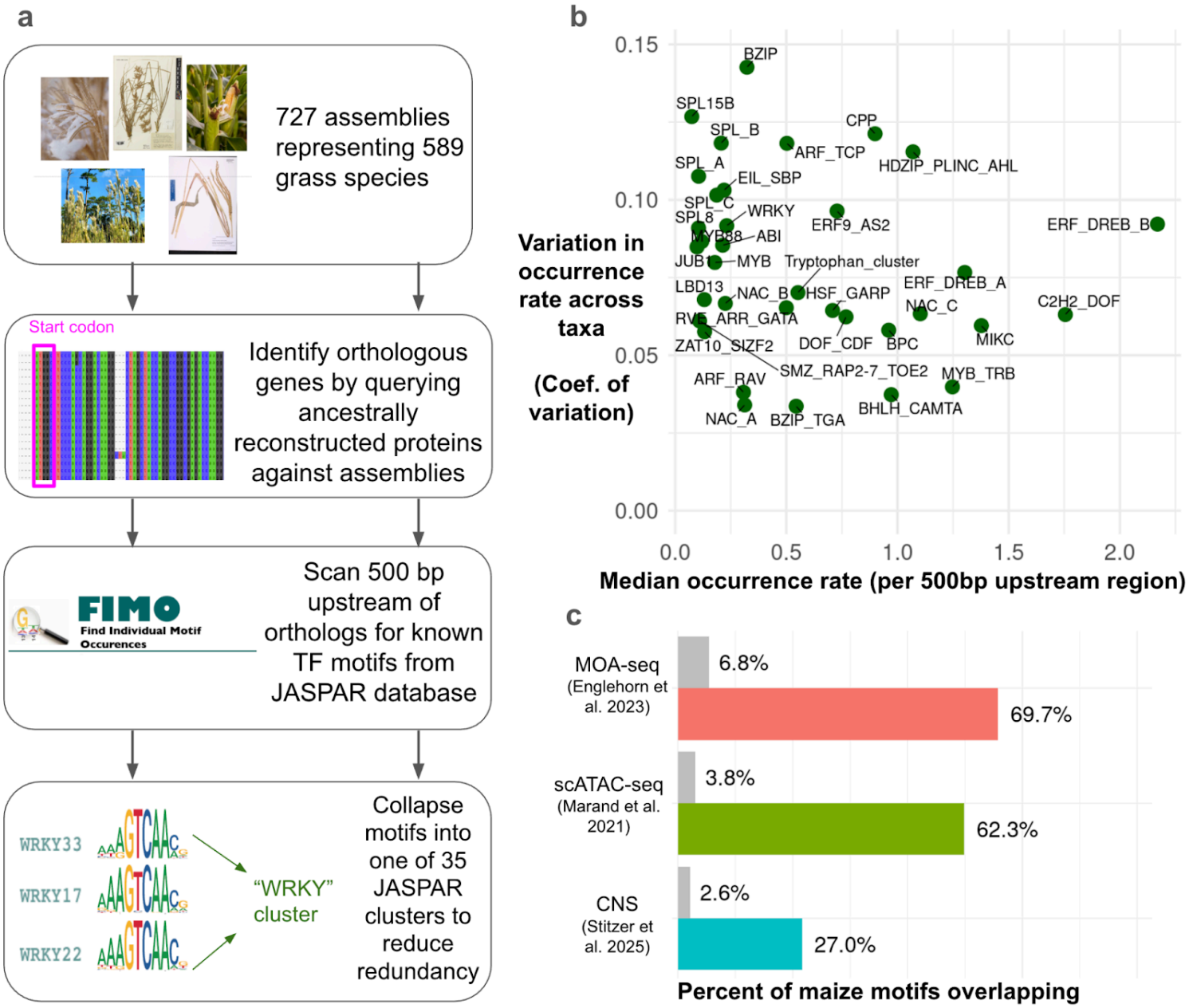

Figure 2. Summary of TF motif profiling across 727 assemblies. (a) Depiction of bioinformatic pipeline to identify TF motif instances upstream of orthologous genes across taxa. (b) Occurrence rate and variation across species for each motif cluster (c) Overlap of maize upstream motifs with known *cis*-regulatory features. Colored bars show the percentage of motifs overlapping intervals from each feature type, while grey bars represent the mean overlap percentage with randomly shuffled intervals.

Motif conservation decays nonlinearly with increasing evolutionary divergence

To track the evolutionary gain and loss of motif instances in grasses, we measured motif conservation by comparing the number of motifs found at maize orthologs with those found in orthologs of the other 588 species (see Methods). The decay in motif conservation relative to maize was extensive. Roughly 60% of maize motif instances were shared with *Sorghum* (~15 million years diverged) at orthologous upstream regions, and 50% were shared with rice across ~80 million years of evolution (Fig. 3a). On average, motif conservation between orthologs did not decay to the level observed between random maize genes, approaching an asymptote of 48% conservation (Fig. 3a). Decay curves were similar when defining motif conservation relative to *Sorghum* (Fig. S6a) and rice (Fig. S6b) instead of maize. Additionally, motifs that co-localized with maize ChIP-seq peaks showed minimal evidence of elevated conservation (Fig. S7), indicating that our motif conservation estimates were not substantially biased by the large number of motifs lacking *in vivo* binding support.

Conservation of motif instances varied widely by gene, with most orthogroups approaching 40-60% conservation between the most distantly related grasses (Fig 3a). Coding sequence conservation explained minimal variance in motif conservation at individual orthogroups (R-squared = 0.03) (Fig. S8). A small number of highly conserved proteins, such as ribosomal and plastid-encoded proteins, exhibited very high rates of motif conservation. However, the motifs detected near plastid genes are likely artifacts of motif scanning, as plastid genes are not regulated by nuclear transcription factors. Orthogroups with higher motif conservation were significantly enriched (Fisher's exact test, $p < 0.05$) for 121 gene ontology (GO) terms. Many of the strongest enrichments were related to regulatory function by transcription factors (Fig. 3b, Supplementary Table 3). Specifically, TF genes (n = 1201)

exhibited slightly higher motif conservation than non-TFs (n = 11,537) (TF median = 57%, non-TF median = 54%; Kolmogorov-Smirnov test, p = 1.7e-14) (Fig. 3c). The difference between TFs and non-TFs persisted when comparing motif conservation between maize and sorghum at syntenic orthologs only (Fig. S9).

In addition to regulatory functions, we observed many significant GO terms among high-conservation genes related to signal transduction (e.g. “calmodulin binding”, “protein phosphorylation”, and “phosphorelay signal transduction system”), cytoskeleton (e.g. “actin filament organization”, “microtubule cytoskeleton organization”, and “cortical microtubule organization”), and development (“flower development”, “abaxial cell fate specification”, and “positive regulation of long-day photoperiodism, flowering”). Low-conservation genes were enriched for 110 terms including “defense response”, “plant-type cell wall”, and “RNA modification” (Fig. S10, Supplementary Table 3). GO terms that were not significantly enriched among either high or low-conservation genes included “response to abiotic stimulus”, “photosynthesis”, and “reproduction”.

Convergent motif gain/loss occurs at hundreds of genes during environmental adaptation

To investigate which motif instances are linked to environmental adaptation, we employed phylogenetic mixed models, which can be used to control for phylogenetic relatedness while associating genomic features with traits across species (Housworth et al. 2004). We characterized each species’ ecological niche as described in (Hsu et al. 2024), summarizing into ten environmental principal coordinate (envPC) axes that together explain 75% of total ecological variability among the natural environments of grasses across the globe.

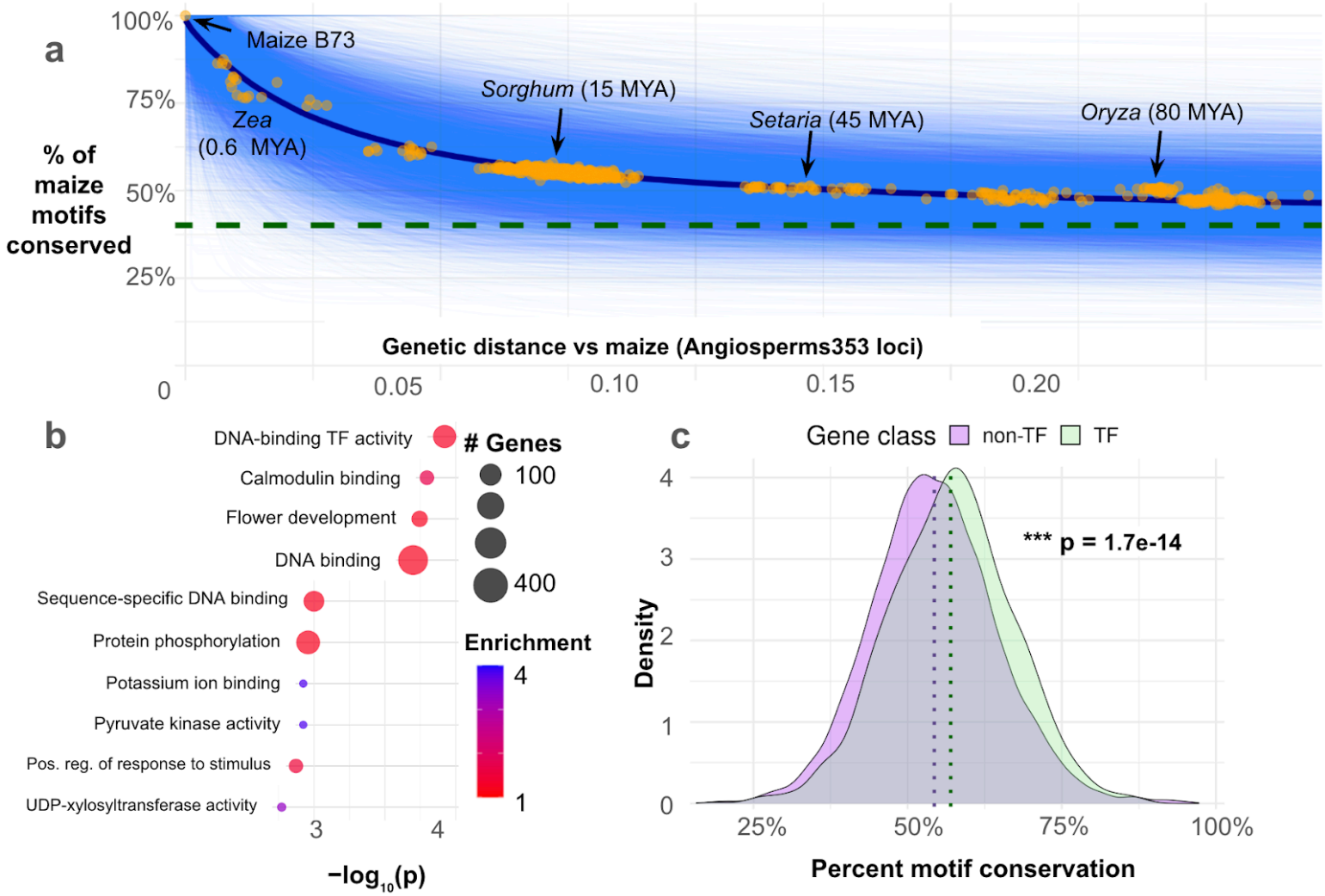

Figure 3. Motif conservation decays nonlinearly and heterogeneously across gene classes. **a.** Percentage of maize motif instances conserved across 589 Poaceae species. Genetic distance was estimated using pairwise distances between maize and Poaceae species at the Angiosperms353 loci. Blue lines represent exponential decay curves fit for 12,784 orthogroups. Orange points show the mean percentage of maize motifs conserved in each Poaceae species across all orthogroups, with an exponential decay curve depicted in dark blue. Dashed green line represents the mean percentage of motifs conserved across 100,000 random pairs of maize genes. Approximate divergence times from maize are shown for key taxa.(Chen et al. 2022; Gallaher et al. 2022)**b.** Enriched gene ontology terms for the orthogroups in the top quartile of motif conservation. **c.** Motif conservation at transcription factor orthogroups (n = 1,204) versus at non-transcription factor orthogroups (n = 11,580) . Median values for each class are shown by dotted vertical lines. P-value is from an asymptotic two-sample Kolmogorov-Smirnov test.

We hypothesized that repeated gains or losses of particular motifs underlie environmental adaptation in grasses. We first tested whether global occurrence rates of motifs across taxa are associated with environmental adaptation, using the envPCs as proxies of environmental diversity (Fig. 4a). Since variable genome sizes have been hypothesized to alter the spacing of regulatory regions (Schmitz et al. 2021), we controlled for the overall density of all motifs as well as the background mono and di-nucleotide content in genomic background regions, which could influence the probability of detecting a motif by chance. Occurrence rates of C2H2-type zinc finger (C2H2/DOF), DNA binding with one finger (DOF/CDF), NAM, ATAF and CUC (NAC), and homodomain-leucine zipper / Plant Zinc Finger / AT-HOOK MOTIF NUCLEAR LOCALIZED (HDZIP/PLINC/AHL) motifs were significantly associated with envPC1 (FDR-adjusted Wald test, $p < 0.01$) (Fig. 4b, Supplementary Table 4), which primarily captures thermal variability across taxa. Additionally, the occurrence rate of basic leucine zipper (BZIP) motifs was significantly associated with envPC2, which captures variation in moisture availability.

We further hypothesized that repeated gains or losses of motif instances at particular orthogroups are associated with environmental adaptation. For each motif-orthogroup combination we associated the number of motifs per assembly at an orthogroup with the ten environmental PCs after controlling for phylogenetic relatedness (Fig. 4c). Across 498,287 models, we scanned for signals of convergent motif gain or loss. We identified 810 unique motif-orthogroup combinations (1025 total) with strongly significant (FDR-adjusted Wald test, $p < 0.01$) environmental associations (Fig. 4d, , Supplementary Table 5), suggesting convergent motif gain or loss across independent lineages.

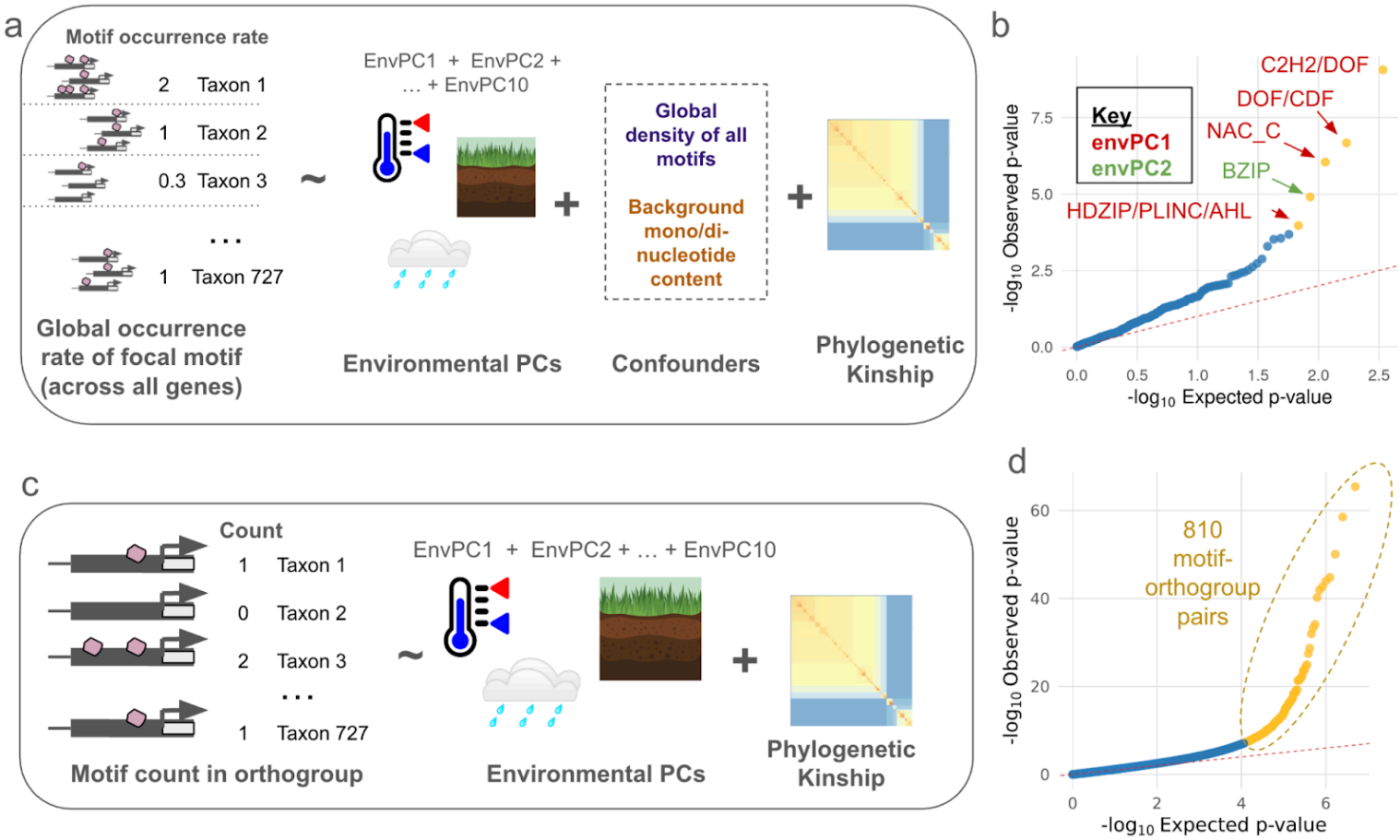

Figure 4. Polygenic turnover at hundreds of motif instances is associated with environmental adaptation. **a.** Setup of phylogenetic mixed models. One model was run for each motif-orthogroup combination for a total of 498,287 models, fitted with ten environmental principle coordinates as fixed effects and a phylogenetic kinship matrix as random effect. **b.** Quantile-quantile plot showing observed p-values from phylogenetic mixed models versus p-values expected under the null hypothesis. Observed p-values were calculated from Wald tests across ~5 million fixed-effect terms. Terms with FDR-corrected p-values < 0.01 are shown in yellow.

Gain of HSF/GARP motifs at an Alpha-N-acetylglucosaminidase gene predicts cold adaptation

We reasoned that the orthogroups most strongly associated with envPCs would be enriched for pathways and processes linked to abiotic stress responses. Indeed, a GO enrichment analysis of the 810 significant orthogroups identified 47 enriched terms, many of which were related to known abiotic stress response processes such as oxidative stress response ("cellular oxidant

306 detoxification", "response to oxidative stress", "oxidoreductase complex", "oxidoreductase
307 activity") and degradation of misfolded proteins ("ubiquitin-dependent ERAD pathway",
308 "proteolysis involved in protein catabolic process", "ubiquitin-specific protease binding",
309 "endoplasmic reticulum unfolded protein response", "threonine-type endopeptidase activity")
310 (Fig. 5a, Supplementary Table 6).

311 We investigated some of the individual motif-orthogroup combinations related to these
312 enriched processes. One of the motif-orthogroup combinations most strongly associated with
313 envPC1 was a Heat Shock Factor / Golden2, ARR-B, Psr1 (HSF/GARP) motif at OG0018131
314 encoding an Alpha-N-acetylglucosaminidase family protein. Alpha-N-acetylglucosaminidase is
315 involved in N-glycan degradation (Ronceret et al. 2008), which has been connected to protein
316 misfolding response under chilling in Arabidopsis (Ma et al. 2016) and alfalfa (Xu et al. 2024)).
317 Additionally, HSF TFs are strongly implicated in thermal stress responses (Guo et al. 2016), and
318 in grasses various HSFs have been linked to misfolded protein responses (Cheng et al. 2015) and
319 chilling tolerance (Gao et al. 2024).

320 Increasing counts of HSF/GARP motifs at OG0018131 predicted lower envPC1 values
321 (Fig. 5b), with the association primarily driven by taxa containing three or more HSF/GARP
322 instances. At least seven independent lineages had three or more copies of the motif (Fig. 5c),
323 representing frost-tolerant genera including *Zizania* (wild rice), *Phragmites*, and *Poa*. Notably,
324 only a fraction of independent cold adaptation events (very roughly 25%) were associated with
325 expansions of this motif, indicating that the strength of convergence at this gene was relatively
326 weak despite strong statistical support (Wald test, p =1.89e-10). Rice contains multiple copies of
327 the motif in an open chromatin region, and its cold-tolerant relative *Zizania palustris* contains
328 additional copies (Fig. 5d).

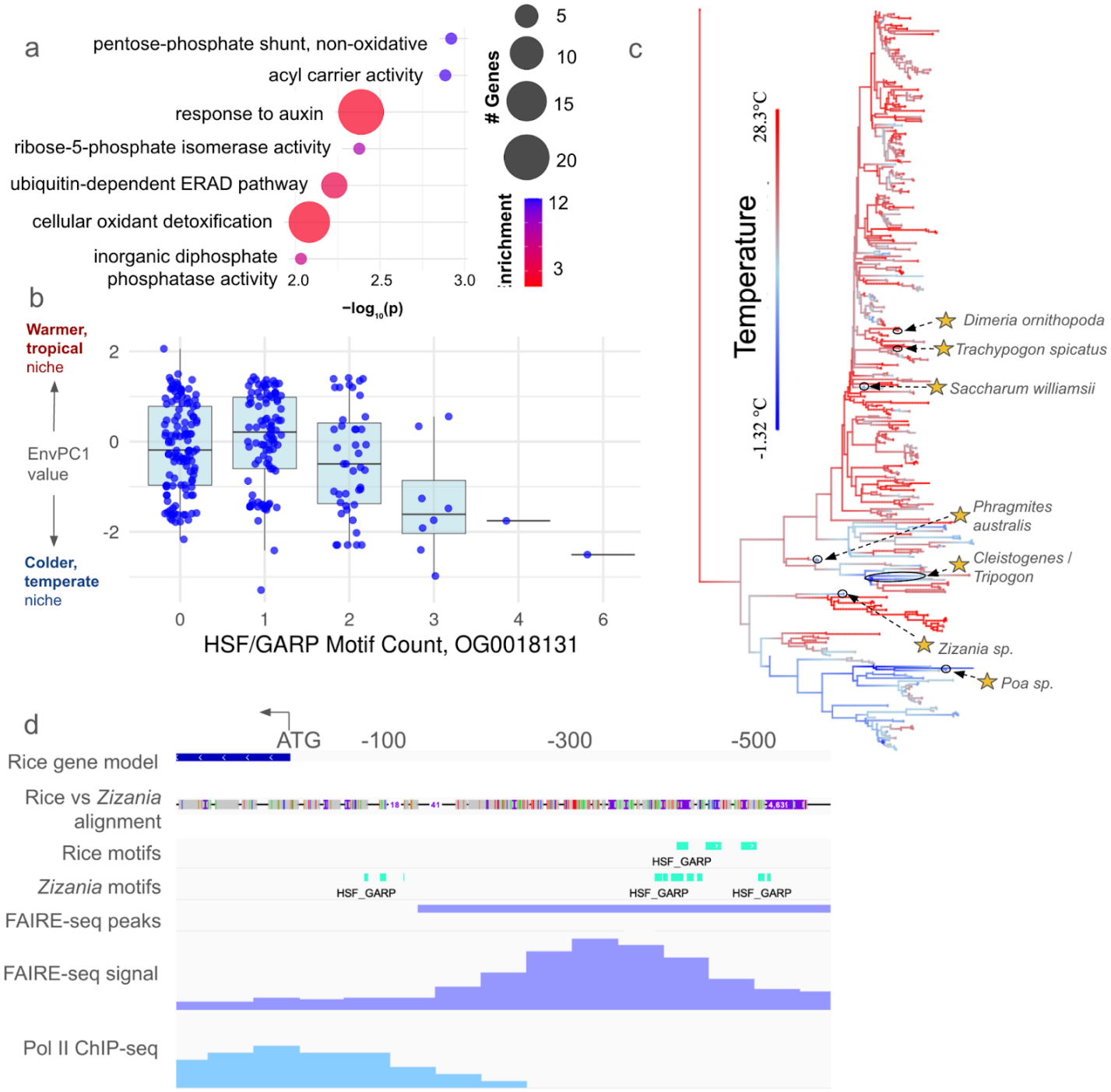

Figure 5. Exploration of top motif-orthogroup pairs. **a.** GO enrichment analysis for 810 significantly-associated orthogroups. **b.** Environmental PC1 values for taxa with varying numbers of HSF/GARP motifs at OG0018131. **c.** Repeated gain and loss of HSF/GARP motifs at OG0018131 across the grass family. Taxa with three or more copies of HSF/GARP are starred. **d.** HSF/GARP motif variation between *Z. palustris* and *O. sativa* at an open chromatin region upstream of OG0018131.

DISCUSSION

Our results support a model of highly polygenic *cis*-regulatory evolution that leverages a conserved regulatory code. The high overlap of UMR-enriched motifs across species suggests that minimal lineage-specific loss of binding preferences has occurred in grasses, consistent with the deep conservation of DNA binding domains observed within many transcription factor families (Yamasaki et al. 2008; Weirauch and Hughes 2011; de Mendoza et al. 2013). Importantly, our analysis does not rule out lineage-specific gains of binding site preference (perhaps via diversification of transcription factor families), which could be investigated using *de novo* motif characterization. Quantitative estimates of binding affinity would be needed to more sensitively detect subtle shifts in binding preferences. Regardless, the high conservation of motif sequences we observed across distantly related species highlights that TFs in grasses share similar binding preferences, likely with transferable regulatory functions across species. Our implicit assumption is that UMR enrichment is evidence of motif function, although motif depletion can also be meaningful since selection on *cis*-regulatory regions may act to inhibit particular TF binding events (He et al. 2011).

While much of grasses' regulatory code appears to have been conserved across a hundred million years of evolution, our work demonstrates that extensive remodeling of individual *cis*-regulatory regions has occurred. Although some TF binding site instances have been shown to be as strongly conserved as coding sequences (Heyndrickx et al. 2014), just 50% of the motif instances we observed were conserved between maize and rice, which are roughly as genetically divergent as humans and mice (Gallaher et al. 2022). Despite extensive turnover overall, conservation of some motif instances appears to persist over deep evolutionary time scales. Motif conservation at orthologous regions decays nonlinearly and tends to approach a minimum level

358 of conservation between 40-60% for most grass orthogroups. This suggests that regulatory
359 regions may often contain a handful of strongly-conserved motif instances alongside many
360 motifs that are rapidly turned over.

361 Our finding that regulatory genes, and specifically transcription factors, have slightly
362 higher motif conservation suggests that a large amount of cis-regulatory evolution within
363 regulatory networks occurs at terminal target genes rather than just at a handful of key regulators.
364 This finding supports a model in which transcription factors and other highly pleiotropic genes
365 exhibit greater constraint, in line with past theoretical and empirical work (Prud'homme et al.
366 2007; Chesmore et al. 2016; Nocchi et al. 2024), but see (Khaipho-Burch et al. 2023). Future
367 studies could explicitly test whether the degree of pleiotropy of a gene predicts conservation of
368 its TF binding sites. Alternatively, background selection on the target gene could be driving
369 conservation of proximal TFBS rather than selection on the TFBS itself, though the low
370 correspondence between coding sequence and motif conservation suggests that this effect is
371 minimal.

372 Our environmental association models suggest that *cis*-regulatory changes in grasses are
373 not concentrated at a handful of key adaptation genes, but occur pervasively across a wide set of
374 diversely functioning genes. We observed weakly convergent motif gain and loss at hundreds of
375 genes with diverse molecular functions, underscoring that environmental adaptation involves
376 altering regulation of a highly diverse set of genes, not just modifying a handful of pathways or
377 processes. These findings are consistent with theoretical findings that suggest that global
378 adaptation of complex traits tends to be associated with many small-effect regulatory changes
379 (Prud'homme et al. 2007). This challenges attempts to leverage cross-species insights for crop
380 genetic engineering. Still, characterization and editing of key TF binding sites remains a

381 promising approach for precisely tuning expression levels.

382 Our results suggest that the phylogenetic mixed model may be an effective approach to
383 nominate key candidate loci using large-scale comparative genomics. With more and more
384 sequenced genomes available, the ability of such comparative genomic approaches to detect
385 convergently evolving genome features will improve (Smith et al. 2020). Future studies could
386 build on this framework by pairing large-scale comparative genomics screens with detailed
387 molecular characterization of the top candidate loci. For example, a variety of mechanisms could
388 underlie the HSF/GARP motif variation we observed at OG0018131. Gains of HSF/GARP
389 motifs at this locus could alter TF binding affinity of HSF/GARP TFs to the region, or perhaps
390 permit combinatorial binding of multiple distinct HSF/GARP TFs that recognize similar motifs.
391 Further work could validate the functional effect of this motif and its relationship with
392 environmental adaptation.

393 A number of tradeoffs were required to scale up our analyses across hundreds of complex
394 plant genomes. The relatively high overlap of gene-proximal motifs intersecting open chromatin
395 and TF binding regions suggests that many gene-proximal motifs are likely to be bound by
396 transcription factors *in vivo*. Still, a significant fraction of the motif instances we characterized
397 are likely to be non-functional. Additionally, studies have found that only a fraction of TF
398 binding changes are due to motif variants (Reddy et al. 2012; Krieger et al. 2022), underscoring
399 that TF binding assays, though far less scalable, currently offer more reliable estimates of TFBS
400 than sequence-based prediction.

401 The low overlap we observed with conserved non-coding regions supports previous
402 findings that alignment-based approaches, while useful for identifying high-importance regions
403 from DNA sequence alone, can miss functional regulatory features (Yang et al. 2015; Kaplow et

404 al. 2023). Alignment-free approaches such as deep learning models hold promise going forward
405 for scalable characterization of *cis*-regulatory regions. The ability of such models to represent the
406 orientation, positioning and co-binding context of *cis*-regulatory elements can offer a more
407 nuanced picture of *cis*-regulatory evolution beyond simple gain and loss of motifs.

408 Our study focused on characterizing motif variants in proximal upstream regions due to
409 limited assembly contiguity. However, changes to distal *cis*-regulatory elements (Clark et al.
410 2006) or *trans*-acting factors also contribute strongly to regulatory adaptation and were not
411 considered in our analyses of motif gain and loss. Another limitation of our approach is that
412 many duplicated gene copies, particularly in the numerous polyploid taxa represented in the
413 dataset, were filtered out. Additionally, because we relied on orthogroups across wide
414 phylogenetic distances to make comparisons, lineage-specific genes that may be key to
415 adaptation were not considered. Similarly, the reliance on convergent evolution in our
416 association models approach may miss adaptive mechanisms that are influenced by evolutionary
417 contingencies and constraints; e.g. lineages using C4 photosynthesis may have different
418 physiological strategies "available" to them than C3 lineages. Future comparative genomic
419 analyses will benefit from complementing broad phylogenetic comparisons with comparisons of
420 closely related species that largely share the same physiology and genetics. Moreover, integrating
421 more detailed estimates of ecological context into population genetic analyses of key taxa would
422 offer greater resolution and nuance beyond our broad species-level estimates of adaptation.

423

424 **CONCLUSION**

425 We performed large-scale comparative genomic analyses across 589 grass species to investigate
426 the *cis*-regulatory basis of diversification and environmental adaptation in grasses. We found that

427 grasses share a deeply conserved regulatory code, with 377 *cis*-regulatory sequence motifs
428 conserved across diverse species. We documented extensive gain and loss of specific motif
429 instances across the grass family, suggesting extensive reorganization of orthologous
430 *cis*-regulatory regions. *Cis*-regulatory changes do not appear to be highly concentrated at
431 particular classes of genes, though they occur slightly less frequently at regulatory genes such as
432 transcription factors. 810 motif-orthogroup combinations show evidence of convergent gain and
433 loss of motifs associated with environmental adaptation. However, even the most strongly
434 significant motifs are only weakly repeatable, underscoring the diversity of *cis*-regulatory routes
435 to environmental adaptation.
436

437 **MATERIALS & METHODS**

438 *Transcription factor motif scanning:* We downloaded 805 TF motifs and clusters from the 2024
439 JASPAR CORE non-redundant plant collection (Rauluseviciute et al. 2023). This set of motifs
440 has strong experimental evidence from TF binding assays such as DAP-seq and ChIP-seq
441 primarily performed in *Arabidopsis*. All motif scans were performed using FIMO from the
442 MEME suite v5.5.7 (Grant et al. 2011) with the parameters --max-strand --no-qvalue
443 --skip-matched-sequence --max-stored-scores 100000000. Homogenous background mono- and
444 di-nucleotide frequencies (=0.25 per nucleotide) were used to avoid species-specific biases in
445 motif detection thresholds. 704 of the 805 motifs in the JASPAR collection were detectable using
446 our parameters, with the rest containing insufficient information content to enable statistically
447 significant matches with FIMO.
448

449 *UMR enrichment analysis:* Unmethylated regions (UMRs) for five diverse grass species
450 (*Sorghum bicolor, Oryza sativa, Brachypodium distachyon, Zea mays, Hordeum vulgare*) were
451 downloaded from (Crisp et al. 2020). Background sequences were generated by di-nucleotide
452 shuffling each UMR region 100 times, preserving local sequence composition. Motif scanning
453 was performed using FIMO on both UMR and background sequences. The fitdist function in R
454 package "fitdistrplus" (Delignette-Muller and Dutang 2015) was used to fit probability
455 distributions to background motif counts. Two-tailed p-values were then calculated to identify
456 significantly over- and under-represented motifs in UMR regions. We quantified how many
457 motifs were significantly over-represented (FDR-adjusted p value < 0.01 from Fisher's exact
458 test) within and between species. The 377 commonly-enriched motifs were used for subsequent
459 analyses.
460 For accessible chromatin enrichment, we downloaded bulk accessible chromatin regions
461 for *S. bicolor, O. sativa, B. distachyon, Z. mays, H. vulgare, Setaria viridis*, and *Arabidopsis*
462 *thaliana* from (Lu et al. 2019) and performed enrichments as described above.
463
464 *Compiling genomic dataset:* We compiled a dataset of 727 genome assemblies representing 589
465 diverse grass species by combining 217 publicly-available assemblies from NCBI (Sayers et al.
466 2024), Phytozome (Goodstein et al. 2012), and CoGE (Lyons and Freeling 2008) with 33
467 long-read assemblies from (Stitzer et al. 2025) and 550 additional assemblies from short reads
468 using the high-throughput assembly pipeline described in (Schulz et al. 2023) (Supplementary
469 Table 2). For this study, we assembled 57 new short-read assemblies from WGS raw reads
470 deposited in the NCBI SRA database (Sayers et al. 2024). We downloaded WGS data from all
471 Poaceae species lacking an existing genome assembly if at least 15GB of WGS data was

available for that species. We then assembled the SRA genomes *de novo* using Megahit v1.2.9 (Li et al. 2015) with minimum kmer size = 31 and default parameters for other assemblies, as described in (Schulz et al. 2023). If < 30GB WGS data was available for the largest WGS accession for a species, we merged WGS data from multiple accessions to improve assembly completeness. We performed three QC steps on our SRA assemblies. First, we ran Kraken2 v2.1.3 (Lu et al. 2022) on raw reads (subsampled to a depth of 10M reads) using the PlusPFP database to verify sample identify on a rough taxonomic scale and flag accessions with high levels of bacterial or fungal contamination. Second, we visualized the assemblies on a *matK* phylogenetic tree against labeled *matK* sequences from the BOLD database (Ratnasingham et al. 2024) to rule out obviously-mislabeled accessions. To generate *matK* alignments, we downloaded a complete *matK* CDS sequence for *Streptochaeta angustifolia* (Genbank: AF164382.1) (Hilu et al. 1999). We queried the *S. angustifolia* sequence against all WGS short read assemblies using minimap2 v2.17 (Li 2018) with the parameters -ax asm20 –eqx -I 100g –secondary=no, extracting the primary alignment in each assembly. Using the extracted *matK* sequences from our WGS assemblies and 7,338 Poaceae *matK* sequences from the BOLD database, we performed multiple alignment with mafft v7.520 (Katoh et al. 2002) --auto. Then we constructed a phylogenetic tree using raxml v8.2.13 (Stamatakis 2014) with the GTRGAMMA model. As a final assessment of assembly quality, we quantified assessed assembly completeness using TABASCO, a BUSCO-like assembly metric designed specifically for use with grasses (Schulz et al. 2023) which labels a set of 5592 query genes as “complete”, “duplicated”, “fragmented”, or “missing”.

Orthogroup construction: Across the 727 genome assemblies, we selected 32 representative high quality long read assemblies to construct orthogroups. In order to avoid potential annotation biases, we ran Helixer (Stiehler et al. 2021) to annotate each of the representative genomes and extracted the protein sequences. Based on protein sequence homology, orthogroups were constructed using OrthoFinder v2.6.4 (Emms and Kelly 2019). In total, we obtained 22,503 orthogroups with homologous sequences represented by more than eight of the representative genomes. From the multiple sequence alignment of each orthgroup, we reconstructed the ancestral protein sequence using the R/phangorn package (Schliep 2011). The ancestral sequences of the orthogroups were used to query the orthologs in all 727 genomes by miniProt v0.13.0 (Li 2023).

Phylogenetic characterization: We calculated phylogenetic relatedness among the 727 studied genomes using the Angiosperms353 loci (McDonnell et al. 2021). We identified the orthogroups homologous to the Angiosperms353 loci using miniProt (v0.13.0) and generated gene trees for each of the orthogroups using RAxML v8.12 (GAMMA + GTR). ASTRAL-Pro v2 (Zhang and Mirarab 2022) was used to reconcile the species tree based on the gene trees. We generated a matrix of shared branch length among any pair of tips in the species tree to represent the relatedness between them. This matrix (hereafter phyloK matrix) is included in our phylogenetic mixed model to account for shared macro-evolutionary history among species. Using the concatenated alignment of the Angiosperms353 loci, we calculated pairwise genetic distance among the 727 taxa based on the K81 model implemented in R/ape::dist.dna() function. We obtained divergence time estimates for *Zea-Tripsacum* from (Chen et al. 2022), and estimated divergence time across Poaceae from (Gallaher et al. 2022).

517 *Motif annotation in orthologous upstream regions:* Alignments were filtered to retain primary
518 alignments that did not contain a frameshift or a premature stop codon, and that began with a
519 start codon. To approximate 5'UTR and promoter regions across assemblies, 500 bp sequence
520 was extracted upstream of the aligned start codon. 500 bp was chosen as a conservative estimate
521 of 5' UTR and promoter regions to maximize species representation in the dataset given the
522 limited contig lengths for the short-read assemblies. If 500 bp upstream sequence was not
523 available, or if the sequence contained > 5% Ns, the sequence was discarded from analysis.
524 Orthogroups with fewer than 200 assemblies represented were dropped. The median number of
525 orthogroups per assembly was 465. To minimize redundant, overlapping motif annotations, we
526 collapsed similar overlapping motifs into a single interval.We defined motif similarity based on
527 membership within the same matrix cluster, as designated by JASPAR 2024 (Rauluseviciute et
528 al. 2023) (Table S1). The 377 motifs with conserved UMR enrichment represented 35 unique
529 clusters. We manually annotated each of the 35 clusters with a descriptor of the motifs contained
530 within (e.g. "HSF/GARP" if the cluster contained motifs from HSF and GARP TFs.) After
531 collapsing overlapping motif instances by cluster, we quantified the number of cluster instances
532 per upstream region for each assembly.
533
534 *Cis-regulatory feature overlap*: We quantified overlap of scATAC-seq peaks, MOA-seq peaks,
535 and conserved non-coding sequences with 278,610 motif instances spanning 500 bp upstream
536 regions in maize B73 v5. Merged scATAC peaks for all maize cell types were downloaded from
537 https://github.com/plantformatics/maize_single_cell_cis_regulatory_atlas/raw/refs/heads/master/
538 all_ACRs.celltype_calls.overlap.bed.gz (Marand et al. 2021) and uplifted to maize B73 v5
539 coordinates using CrossMap v.0.7.3 (Zhao et al. 2014) and a chain file downloaded from

540 MaizeGDB. 168,009 peaks were successfully lifted over and used for the overlap analysis.
541 Additionally, 701,701 merged MOA-seq peaks for 21 maize hybrids projected onto B73
542 coordinates were obtained from (Engelhorn et al. 2023) and 1,664,343 conserved non-coding
543 sequences for the Andropogoneae clade were obtained from (Stitzer et al. 2025). We used the
544 intersect function from bedtools v2.29.2 (Quinlan and Hall 2010) with -u to quantify how many
545 motifs overlapped each feature type. Null motif overlap values were calculated using 100
546 shuffled intervals for each feature type, using bedtools shuffle with -noOverlapping to prevent
547 shuffled intervals from overlapping
548
549 *Motif abundance and variability:* We used our motif counts in 500 bp upstream regions to
550 estimate 1) the median occurrence rate and 2) coefficient of variation of each motif cluster across
551 assemblies. To provide more stable estimates, we filtered out assemblies containing fewer than
552 1000 orthogroups. We calculated occurrence rates of each motif for each assembly by dividing
553 the total number of instances of a particular motif in 500 bp upstream regions by the number of
554 500 bp upstream regions represented in that assembly, effectively calculating the average number
555 of motif instances per upstream region. We then calculated the median and coefficient of
556 variation for motif occurrence rates across all assemblies.
557
558 *Conservation of motif instances:* Motif conservation was defined for each assembly relative to
559 Maize B73, with the percent conservation of maize motifs at a single orthogroup defined as 1 - (#
560 motifs present at the target assembly gene / # of motifs present at the maize ortholog). As with
561 most estimates of TFBS turnover, we did not consider motifs present in other lineages but absent

562 in maize. For example, conservation at a single ortholog with three motif types present might be
563 quantified as follows:

	Motif A instances	Motif B instances	Motif C instances	Total maize motifs present
Target assembly	1	0	2	2
Maize	1	1	1	3
% conservation	100%	0%	100%	67%

564

565 We calculated motif conservation relative to maize both at individual orthogroups ("local") and
566 across all 12,784 orthogroups present in maize ("global"). For global estimates of motif
567 conservation, we excluded assemblies with < 1000 orthogroups shared with maize. For local and
568 global conservation, we plotted motif conservation against genetic distance (calculated using the
569 Angiosperms353 loci as described above). We fit exponential decay curves of the form
570 $y = a * \exp(-bx) + c$ to the relationship between genetic distance (x) and motif conservation (y).
571 To identify functional enrichments for orthogroups with high and low motif conservation,
572 we calculated the mean percent conservation across taxa for maize motifs. Using the mean
573 conservation values for each orthogroup, we identified orthogroups within the top and bottom
574 quartiles of motif conservation. The top and bottom quartiles corresponded to 3051 and 3031
575 maize genes, which we mapped to orthogroups using miniProt (v0.13.0) between the maize
576 sequences and the reconstructed ancestral sequences of the orthogroups in this study. We then
577 performed GO enrichment using the R/topGO package (Alexa and Rahnenführer 2009) with the
578 "weight01" algorithm, which considers hierarchical relationships among terms. For background
579 genes, we used the remaining ~9700 orthogroups not contained in the target quartile. Using
580 maize B73 v5 GO annotations downloaded from MaizeGDB (Woodhouse et al. 2021), we tested

581 for significant cellular component, biological process, and molecular function terms and
582 calculated p values from Fisher statistics.

583 To compare motif conservation at TF vs non-TF genes, we downloaded a list of maize
584 TFs from Grassius (Gavgani et al. 2023) corresponding to 1,204 distinct orthogroups. Using
585 orthogroup-level measures of motif conservation, we used an asymptotic two-sample
586 Kolmogorov-Smirnov test to evaluate the difference in conservation between TFs and non-TFs
587 (all orthogroups not contained in the Grassius list).

588 To compare motif conservation at genes with ChIP-seq support, we downloaded
589 ChIP-seq data from four maize TFs (Tu et al. 2020) matched to motifs used in our analyses:
590 ereb17 (MA1818.1), glk53 (MA1830.2), bhlh47 (MA1834.2), and hb34 (MA1824.2). For each
591 TF, we identified a set of orthogroups for which the 500 bp upstream region of the maize
592 ortholog intersected with a ChIP-seq peak. Then, for the motif clusters corresponding to each
593 ChIP-seq TF we quantified conservation of motif instances between maize and every assembly
594 across 1) all orthogroups 2) all orthogroups intersecting a maize ChIP-seq peak.

595 *Environmental characterization:* We followed an environmental characterization pipeline
596 previously described by (Hsu et al. 2024) to characterize the environmental niches of the studied
597 species. Briefly, for each species, we retrieved the geographical coordinates of occurrence from
598 BIEN (Maitner et al 2017) and GBIF databases and obtained various environmental features in
599 which the species occurs. We were able to characterize the habitat environments for 706 out of
600 727 taxa and thus excluded taxa where we were unable to curate environmental data from
601 subsequent analysis (Hsu 2025). The variation in habitat environments amongst the diverse
602 species were summarized into environmental principal components (“envPCs”). Each envPC
603 captures a different combination of spatial and climatic patterns, with the multivariate

604 decomposition process resulting in synthetic variables (the envPCs) that preserve important
605 statistical properties, such as orthogonality, lack of autocorrelation, and normality. The top 10
606 environmental PCs were modeled as predictor variables in our phylogenetic mixed models.
607
608 *Cross-species environmental association models:* For models associating global motif
609 occurrence rates with environment, we fit 35 phylogenetic mixed models of the following form
610 with ASReml-R v4.2 (Butler et al. 2009), with one model per motif and one observation per
611 assembly in each model:
612
613 *Focal motif occurrence rate ~ environmental PC1 + environmental PC2 +...+ environmental PC10 + Global motif*
614 *density (all motifs) + Mono/di-nucleotide frequency PCs 1-5 + phyloK*
615
616 The phylogenetic relationship matrix (phyloK) was fit as a random effect to control for shared
617 evolution across species. All other predictor variables were fit as fixed effects. Occurrence rates
618 for each focal motif were calculated as described in the "Motif abundance and variability"
619 section above. We quantified global motif density (across all motifs) for each assembly by
620 summing all 35 individual motif occurrence rates. Mono/di-nucleotide principle coordinates
621 were used as covariates to control for background genome characteristics such as GC content,
622 which influence motif detection rates. To estimate mononucleotide and dinucleotide frequencies
623 in neutral genomic background regions, we extracted sequences for all introns shorter than 150bp
624 using the miniProt alignments previously described in the "Orthogroup construction" section. We
625 used the fasta-get-markov function from the MEME suite with -m 1 to calculate
626 mono/di-nucleotide frequencies, and five principle coordinate axes were generated from these
627 frequencies using the R/ade4 package (Dray and Dufour 2007). We fit the envPCs terms as fixed

effects, along with the mono/di-nucleotide content and global motif density covariate terms. We calculated p values for each envPC term using Wald tests. We plotted the distribution of the 350 observed p values against p values expected under a uniform distribution to identify deviations from the null hypothesis.

For orthogroup-specific models of motif-environment associations, we used ASReml-R v4.2 to run phylogenetic mixed models of the following form, with one observation per assembly:

Motif instances in orthogroup ~ environmental PC1 + environmental PC2 +...+ environmental PC10 + phyloK

Environmental PCs (envPCs) were specified as fixed effects with the phylogenetic relationship matrix (phyloK) as a random effect to control for shared evolution across species. In total 498,287 models were run (# of motifs x # of orthogroups). Models with over 95% variance explained by phyloK were discarded to minimize unstable fixed effect estimates. Wald tests were performed on each envPC term to calculate p values. We plotted the distribution of the ~ 5 million observed p values against p values expected under a uniform distribution to identify deviations from the null hypothesis of no association between environment and motif occurrence. To visualize how motif gain/loss is associated with temperature adaptation events, we reconstructed ancestral temperature values across the Poaceae phylogeny using the contMap function from phytools v2.3-0 (Revell 2024).

Visual comparison of orthologous regions: To compare orthologous upstream regions, we performed pairwise whole-genome alignments between *O. sativa* var. ZhenShan97 and *Zizania palustris* using Anchorwave v.1.2.2 (Song et al. 2022) with parameters -R 2 -Q 1. We generated chain files from Anchorwave alignments using a custom “MAFtoChain” script, then lifted over

motifs into rice coordinates using CrossMap v0.7.3 (Zhao et al. 2014). To visualize open chromatin regions, we used RiceENCODE to download FAIRE-seq and RNA Polymerase II ChIP-seq tracks from (Zhao et al. 2020).

DATA AVAILABILITY: Short read genome assemblies will be made publicly available on Ag Data Commons upon publication. Code to generate the analyses and figures can be found at https://github.com/maize-genetics/poaceae_tfbs.

ACKNOWLEDGEMENTS:

Funding: This work was supported by NSF PanAnd Grant Award #1822330 and USDA-ARS Project Number 8062-21000-052-004-A. E.S.B. is supported by the USDA-ARS (ARS project number 8062-21000-052-000-D). A.J.S. was supported by NSF GRFP DGE – 2139899. M.C.S. was supported by NSF PRFB 1907343. A.P.M. is supported by NIH NIGMS 1R00GM144742. Computational resources were provided by the SCINet project and the AI Center of Excellence of the USDA Agricultural Research Service, ARS project numbers 0201-88888-003-000D and 0201-88888-002-000D. Additional computational resources and data management were provided by the Bioinformatics Facility (RRID:SCR_021757) at the Cornell Institute of Biotechnology.

Collaborators: We thank Julia Englehorn for early access to MOA-seq data. Thanks to Chelsea Specht, Charles Danko, Jian Hua, and the entirety of the Buckler lab for valuable feedback throughout the project.

Germplasm collection

Obtaining plant material for this project would not be possible without our many collaborators in the field and herbaria. We're grateful for their expertise, willingness to aid in field collecting, and

sharing of precious herbarium material. In no particular order, we thank the following individuals, governmental organizations, and research institutions:

Individuals

Taylor AuBuchon-Elder led the germplasm collection and plant maintenance with help from the following individuals: Donald Danforth Plant Science Center Plant Growth Facility staff, Rémy Pasquet, Cassiano Welker, Chrissy McAllister, Pat Minx, Bess Bookout, Maria Vorontsova, Jordan Teisher, Pete Lowry, Sarah Mathews, Richard Jobson, Tim Teetaert, Julie Pelc, Rev. Dennis Testerman, Russel Juelg, James Cole, Ron Day, Courtney Angelo, Chris Matson, Douglas Rogers, Michael McKain, Marshall Shaw, Arthur Stiles, Chuck Byrd, Kyle Dillard, Robert Findling, Nancy Sferra, Jonathan Bailey, Lynn Riedel, Brian Anacker, Kirsti Harms, Brandon Crawford, Charlotte Reemts, Matt McCaw, Kevin Thuesen, Michelle Bertelsen, Ryan Middleton, Wesley Newman

Plant vouchers from available material were deposited at Missouri Botanical Garden Herbarium; and for Australian wild collections, duplicates were deposited at Australia's National Herbarium in Canberra.

Herbaria (silica dried leaf tissue and herbarium specimen tissue)

Missouri Botanical Gardens, Royal Botanic Garden Kew, Muséum national d'Histoire naturelle, Australia National Herbarium, Queensland Herbarium, Northern Territory Herbarium, National Herbarium of New South Wales

Governmental organizations (permits, permissions, and live plant or seed material)

Australian National Government, New South Wales National Parks and Wildlife Service, Queensland Parks and Wildlife Service, Queensland Department of Environment Service, Northern Territory Government, Northern Territory Department of Natural Resources, Victoria Department of Environment, Land, Water, and Planning, North Carolina Department of Agriculture and Consumer Services, NC Plant Conservation Program, US National Park Service, US Department of Agriculture, City of Boulder Open Space and Mountain Parks, Florida Park Service and Department of Environmental Protection, City of Austin Water Quality Protection Lands

Non-profits, NGOs, Research institutes (permits, permissions, and live plant material)

Katy Prairie Conservatory, Lady Bird Johnson Wildflower Center, The Nature Conservancy (Alabama, Arizona, Florida, Maine, Manitoba, Missouri, New Mexico, Texas), Native Prairie Association of Texas, University of Alabama, New Jersey Conservation Foundation, University of Puerto Rico at Mayagüez

All collections were done in compliance with the Nagoya Protocol, with permits as required by local authorities.

We acknowledge and honor the many custodians and stewards of wild and domesticated grass diversity worldwide who have shaped the study, preservation, and cultivation of grasses.

References

Alexa A, Rahnenführer J. 2009. Gene set enrichment analysis with topGO. *Bioconductor Improv* 27:1–26.

Andrews G, Fan K, Pratt HE, Phalke N, Zoonomia Consortium§, Karlsson EK, Lindblad-Toh K, Gazal S, Moore JE, Weng Z, et al. 2023. Mammalian evolution of human cis-regulatory elements and transcription factor binding sites. *Science* 380:eabn7930.

Bennett MD, Smith JB. 1976. Nuclear dna amounts in angiosperms. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 274:227–274.

Bennetzen JL, Freeling M. 1993. Grasses as a single genetic system: genome composition, collinearity and compatibility. *Trends Genet.* 9:259–261.

Bouchenak-Khelladi Y, Verboom GA, Savolainen V, Hodkinson TR. 2010. Biogeography of the grasses (Poaceae): a phylogenetic approach to reveal evolutionary history in geographical space and geological time. *Bot. J. Linn. Soc.* 162:543–557.

Buell CR. 2009. Poaceae genomes: going from unattainable to becoming a model clade for comparative plant genomics. *Plant Physiol.* 149:111–116.

Butler D, Cullis B, Gilmour A, Gogel B. 2009. mixed models for S language environments ASReml-R reference manual ASReml estimates variance components under a general linear mixed model by residual maximum likelihood (REML). Available from: https://asreml.kb.vsni.co.uk/wp-content/uploads/sites/3/ASReml-R-Reference-Manual-4.2.pdf

Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* 134:25–36.

Cheng Q, Zhou Y, Liu Z, Zhang L, Song G, Guo Z, Wang W, Qu X, Zhu Y, Yang D. 2015. An alternatively spliced heat shock transcription factor, OsHSFA2dI, functions in the heat stress-induced unfolded protein response in rice. *Plant Biol. (Stuttg.)* 17:419–429.

Chen L, Luo J, Jin M, Yang N, Liu X, Peng Y, Li W, Phillips A, Cameron B, Bernal JS, et al. 2022. Genome sequencing reveals evidence of adaptive variation in the genus Zea. *Nat. Genet.* 54:1736–1745.

Chesmore KN, Bartlett J, Cheng C, Williams SM. 2016. Complex Patterns of Association between Pleiotropy and Transcription Factor Evolution. *Genome Biol. Evol.* 8:3159–3170.

Clark RM, Wagler TN, Quijada P, Doebley J. 2006. A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture. *Nat. Genet.* 38:594–597.

Crisp PA, Marand AP, Noshay JM, Zhou P, Lu Z, Schmitz RJ, Springer NM. 2020. Stable

unmethylated DNA demarcates expressed genes and their cis-regulatory space in plant genomes. *Proc. Natl. Acad. Sci. U. S. A.* 117:23991–24000.

Cui Y, Cao Q, Li Y, He M, Liu X. 2023. Advances in cis-element- and natural variation-mediated transcriptional regulation and applications in gene editing of major crops. *J. Exp. Bot.* 74:5441–5457.

Delignette-Muller ML, Dutang C. 2015. fitdistrplus: AnRPackage for Fitting Distributions. *J. Stat. Softw.* [Internet] 64. Available from: http://dx.doi.org/10.18637/jss.v064.i04

Dray S, Dufour A. 2007. The ade4 package: Implementing the duality diagram for ecologists. *Journal of Statistical Software* 22:1–20.

Edwards EJ, Smith SA. 2010. Phylogenetic analyses reveal the shady history of C_4 grasses. *Proc. Natl. Acad. Sci. U. S. A.* 107:2532–2537.

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20:238.

Engelhorn J, Snodgrass SJ, Kok A, Seetharam AS, Schneider M, Kiwit T, Singh A, Banf M, Khaipho-Burch M, Runcie DE, et al. 2023. Phenotypic variation in maize can be largely explained by genetic variation at transcription factor binding sites. *bioRxiv* [Internet]:2023.08.08.551183. Available from: https://www.biorxiv.org/content/10.1101/2023.08.08.551183v1

Fjellheim S, Boden S, Trevaskis B. 2014. The role of seasonal flowering responses in adaptation of grasses to temperate climates. *Front. Plant Sci.* 5:431.

Gallaher TJ, Peterson PM, Soreng RJ, Zuloaga FO, Li D-Z, Clark LG, Tyrrell CD, Welker CAD, Kellogg EA, Teisher JK. 2022. Grasses through space and time: An overview of the biogeographical and macroevolutionary history of Poaceae. *J. Syst. Evol.* 60:522–569.

Gao L, Pan L, Shi Y, Zeng R, Li M, Li Z, Zhang X, Zhao X, Gong X, Huang W, et al. 2024. Genetic variation in a heat shock transcription factor modulates cold tolerance in maize. *Mol. Plant* 17:1423–1438.

Gavgani HN, Grotewold E, Gray J. 2023. Methodology for constructing a knowledgebase for plant gene regulation information. *Methods Mol. Biol.* 2698:277–300.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, et al. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40:D1178–D1186.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27:1017–1018.

Grass Phylogeny Working Group II. 2012. New grass phylogeny resolves deep evolutionary relationships and discovers C4 origins. *New Phytol.* 193:304–312.

Guo M, Liu J-H, Ma X, Luo D-X, Gong Z-H, Lu M-H. 2016. The plant heat stress transcription factors (HSFs): Structure, regulation, and function in response to abiotic stresses. *Front. Plant Sci.* 7:114.

Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, Forczek E, Joly-Lopez Z, Steffen JG, Hazzouri KM, et al. 2013. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat. Genet.* 45:891–898.

He BZ, Holloway AK, Maerkl SJ, Kreitman M. 2011. Does positive selection drive transcription factor binding site turnover? A test with Drosophila cis-regulatory modules. *PLoS Genet.* 7:e1002053.

Hendelman A, Zebell S, Rodriguez-Leal D, Dukler N, Robitaille G, Wu X, Kostyun J, Tal L, Wang P, Bartlett ME, et al. 2021. Conserved pleiotropy of an ancient plant homeobox gene uncovered by cis-regulatory dissection. *Cell* 184:1724–1739.e16.

Heyndrickx KS, Van de Velde J, Wang C, Weigel D, Vandepoele K. 2014. A functional and evolutionary perspective on transcription factor binding in Arabidopsis thaliana. *Plant Cell* 26:3894–3910.

Hilu KW, Alice LA, Liang H. 1999. Phylogeny of Poaceae inferred from matK sequences. *Ann. Mo. Bot. Gard.* 86:835.

Housworth EA, Martins EP, Lynch M. 2004. The phylogenetic mixed model. *Am. Nat.* 163:84–96.

Hsu SK, Schulz AJ, Hale CO, et al. The Genomic basis of environmental adaptation in Poaceae. Unpublished data.

Hsu S-K. 2025. Occurrence data for 614 Poaceae species with associated genome assemblies. Available from: http://dx.doi.org/10.5281/zenodo.14968186

Hsu S-K, Emmett BD, Haffke AJ, Costa-Neto G, Schulz AJ, Lepak N, La T, AuBuchon-Elder TM, Hale CO, Raglin SS, et al. 2024. Contrasting Rhizosphere Nitrogen Dynamics in Andropogoneae Grasses: Implications for Sustainable Agriculture. *bioRxiv* [Internet]:2024.06.03.597142. Available from: https://www.biorxiv.org/content/10.1101/2024.06.03.597142v1

Hu Z-C, Majda M, Sun H-R, Zhang Y, Ding Y-N, Yuan Q, Su T-B, Lü T-F, Gao F, Xu G-X, et al. 2024. Evolution of a SHOOTMERISTEMLESS transcription factor binding site promotes fruit shape determination. *Nat. Plants*:1–13.

Jayaram N, Usvyat D, R Martin AC. 2016. Evaluating tools for transcription factor binding site prediction. *BMC Bioinformatics* 17:547.

Jiang H, Shi Y, Liu J, Li Z, Fu D, Wu S, Li M, Yang Z, Shi Y, Lai J, et al. 2022. Natural polymorphism of ZmICE1 contributes to amino acid metabolism that impacts cold tolerance in maize. *Nat Plants* 8:1176–1190.

827 Kaplow IM, Lawler AJ, Schäffer DE, Srinivasan C, Sestili HH, Wirthlin ME, Phan BN, Prasad
828 K, Brown AR, Zhang X, et al. 2023. Relating enhancer genetic variation across mammals to
829 complex phenotypes using machine learning. *Science* 380:eabm7993.

830 Katoh K, Misawa K, Kuma K-I, Miyata T. 2002. MAFFT: a novel method for rapid multiple
831 sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30:3059–3066.

832 Kellogg EA. 2015. Flowering Plants. Monocots: Poaceae. 2015th ed. Cham, Switzerland:
833 Springer International Publishing

834 Khaipho-Burch M, Ferebee T, Giri A, Ramstein G, Monier B, Yi E, Romay MC, Buckler ES.
835 2023. Elucidating the patterns of pleiotropy and its biological relevance in maize. *PLoS*
836 *Genet.* 19:e1010664.

837 King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science*
838 188:107–116.

839 Krieger G, Lupo O, Wittkopp P, Barkai N. 2022. Evolution of transcription factor binding
840 through sequence variations and turnover of binding sites. *Genome Res.* 32:1099–1111.

841 Lehmann CER, Griffith DM, Simpson KJ, Michael Anderson T, Archibald S, Beerling DJ, Bond
842 WJ, Denton E, Edwards EJ, Forrestel EJ, et al. Functional diversification enabled grassy
843 biomes to fill global climate space. Available from: http://dx.doi.org/10.1101/583625

844 Lehti-Shiu MD, Panchy N, Wang P, Uygun S, Shiu S-H. 2017. Diversity, expansion, and
845 evolutionary novelty of plant DNA-binding transcription factor families. *Biochim. Biophys.*
846 *Acta Gene Regul. Mech.* 1860:3–20.

847 Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. 2015. MEGAHIT: an ultra-fast single-node
848 solution for large and complex metagenomics assembly via succinct de Bruijn graph.
849 *Bioinformatics* 31:1674–1676.

850 Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*
851 34:3094–3100.

852 Li H. 2023. Protein-to-genome alignment with miniprot. *Bioinformatics* [Internet] 39. Available
853 from: http://dx.doi.org/10.1093/bioinformatics/btad014

854 Lu J, Rincon N, Wood DE, Breitwieser FP, Pockrandt C, Langmead B, Salzberg SL, Steinegger
855 M. 2022. Metagenome analysis using the Kraken software suite. *Nat. Protoc.*
856 17:2815–2839.

857 Lu Z, Marand AP, Ricci WA, Ethridge CL, Zhang X, Schmitz RJ. 2019. The prevalence,
858 evolution and chromatin signatures of plant regulatory elements. *Nat Plants* 5:1250–1259.

859 Lyons E, Freeling M. 2008. How to usefully compare homologous plant genes and chromosomes
860 as DNA sequences: How to usefully compare plant genomes. *Plant J.* 53:661–673.

Mahendrawada L, Warfield L, Donczew R, Hahn S. 2025. Low overlap of transcription factor DNA binding and regulatory targets. *Nature*:1–9.

Ma J, Wang D, She J, Li J, Zhu J-K, She Y-M. 2016. Endoplasmic reticulum-associated N-glycan degradation of cold-upregulated glycoproteins in response to chilling stress in Arabidopsis. *New Phytol.* 212:282–296.

Marand AP, Chen Z, Gallavotti A, Schmitz RJ. 2021. A cis-regulatory atlas in maize at single-cell resolution. *Cell* 184:3041–3055.e21.

Marand AP, Eveland AL, Kaufmann K, Springer NM. 2022. cis-Regulatory Elements in Plant Development, Adaptation, and Evolution. *Annu. Rev. Plant Biol.* [Internet]. Available from: http://dx.doi.org/10.1146/annurev-arplant-070122-030236

Mascher M, Marone MP, Schreiber M, Stein N. 2024. Are cereal grasses a single genetic system? *Nat. Plants* [Internet]. Available from: http://dx.doi.org/10.1038/s41477-024-01674-3

McDonnell AJ, Baker WJ, Dodsworth S, Forest F, Graham SW, Johnson MG, Pokorny L, Tate J, Wicke S, Wickett NJ. 2021. Exploring Angiosperms353: Developing and applying a universal toolkit for flowering plant phylogenomics. *Appl. Plant Sci.* [Internet] 9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8312743/

McSteen P, Kellogg EA. 2022. Molecular, cellular, and developmental foundations of grass diversity. *Science* 377:599–602.

de Mendoza A, Sebé-Pedrós A, Šestak MS, Matejcic M, Torruella G, Domazet-Loso T, Ruiz-Trillo I. 2013. Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. *Proc. Natl. Acad. Sci. U. S. A.* 110:E4858–E4866.

Meng X, Liang Z, Dai X, Zhang Y, Mahboub S, Ngu DW, Roston RL, Schnable JC. 2021. Predicting transcriptional responses to cold stress across plant species. *Proc. Natl. Acad. Sci. U. S. A.* [Internet] 118. Available from: http://dx.doi.org/10.1073/pnas.2026330118

Nitta KR, Jolma A, Yin Y, Morgunova E, Kivioja T, Akhtar J, Hens K, Toivonen J, Deplancke B, Furlong EEM, et al. 2015. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *Elife* [Internet] 4. Available from: http://dx.doi.org/10.7554/eLife.04837

Nocchi G, Whiting JR, Yeaman S. 2024. Repeated global adaptation across plant species. *Proc. Natl. Acad. Sci. U. S. A.* 121:e2406832121.

O'Malley RC, Huang S-SC, Song L, Lewsey MG, Bartlett A, Nery JR, Galli M, Gallavotti A, Ecker JR. 2016. Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell* 165:1280–1292.

Prud'homme B, Gompel N, Carroll SB. 2007. Emerging principles of regulatory evolution. *Proc. Natl. Acad. Sci. U. S. A.* 104 Suppl 1:8605–8612.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842.

Ratnasingham S, Wei C, Chan D, Agda J, Agda J, Ballesteros-Mejia L, Boutou HA, El Bastami ZM, Ma E, Manjunath R, et al. 2024. BOLD v4: A centralized bioinformatics platform for DNA-based biodiversity data. *Methods Mol. Biol.* 2744:403–441.

Rauluseviciute I, Riudavets-Puig R, Blanc-Mathieu R, Castro-Mondragon JA, Ferenc K, Kumar V, Lemma RB, Lucas J, Chèneby J, Baranasic D, et al. 2023. JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* [Internet]. Available from: http://dx.doi.org/10.1093/nar/gkad1059

Reddy TE, Gertz J, Pauli F, Kucera KS, Varley KE, Newberry KM, Marinov GK, Mortazavi A, Williams BA, Song L, et al. 2012. Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Res.* 22:860–869.

Revell LJ. 2024. phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things). *PeerJ* 12:e16505.

Riechmann JL, Heard J, Martin G, Reuber L, Jiang C, Keddie J, Adam L, Pineda O, Ratcliffe OJ, Samaha RR, et al. 2000. Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science* 290:2105–2110.

Rodgers-Melnick E, Vera DL, Bass HW, Buckler ES. 2016. Open chromatin reveals the functional maize genome. *Proc. Natl. Acad. Sci. U. S. A.* 113:E3177–E3184.

Ronceret A, Gadea-Vacas J, Guilleminot J, Devic M. 2008. The alpha-N-acetyl-glucosaminidase gene is transcriptionally activated in male and female gametes prior to fertilization and is essential for seed development in Arabidopsis. *J. Exp. Bot.* 59:3649–3659.

Savadel SD, Hartwig T, Turpin ZM, Vera DL, Lung P-Y, Sui X, Blank M, Frommer WB, Dennis JH, Zhang J, et al. 2021. The native cistrome and sequence motif families of the maize ear. *PLoS Genet.* 17:e1009689.

Sayers EW, Beck J, Bolton EE, Brister JR, Chan J, Comeau DC, Connor R, DiCuccio M, Farrell CM, Feldgarden M, et al. 2024. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 52:D33–D43.

Schliep KP. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics* 27:592–593.

Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S, et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* 328:1036–1040.

Schmitz RJ, Grotewold E, Stam M. 2021. Cis-regulatory sequences in plants: their importance, discovery, and future challenges. *Plant Cell* [Internet]. Available from: https://academic.oup.com/plcell/advance-article-pdf/doi/10.1093/plcell/koab281/41787458/koab281.pdf

Schulz AJ, Hsu SK, Hale CO, et al. The molecular evolution of perenniality across the grasses. Unpublished data.

Schulz AJ, Zhai J, AuBuchon-Elder T, El-Walid M, Ferebee TH, Gilmore EH, Hufford MB, Johnson LC, Kellogg EA, La T, et al. 2023. Fishing for a reelGene: evaluating gene models with evolution and machine learning. *bioRxiv* [Internet]. Available from: https://www.biorxiv.org/content/10.1101/2023.09.19.558246.abstract

Signor SA, Nuzhdin SV. 2018. The Evolution of Gene Expression in cis and trans. *Trends Genet.* 34:532–544.

Smith SD, Pennell MW, Dunn CW, Edwards SV. 2020. Phylogenetics is the New Genetics (for Most of Biodiversity). *Trends Ecol. Evol.* 35:415–425.

Song B, Buckler ES, Wang H, Wu Y, Rees E, Kellogg EA, Gates DJ, Khaipho-Burch M, Bradbury PJ, Ross-Ibarra J, et al. 2021. Conserved noncoding sequences provide insights into regulatory sequence and loss of gene expression in maize. *Genome Res.* [Internet]. Available from: http://dx.doi.org/10.1101/gr.266528.120

Song B, Marco-Sola S, Moreto M, Johnson L, Buckler ES, Stitzer MC. 2022. AnchorWave: Sensitive alignment of genomes with high sequence diversity, extensive structural polymorphism, and whole-genome duplication. *Proc. Natl. Acad. Sci. U. S. A.* [Internet] 119. Available from: http://dx.doi.org/10.1073/pnas.2113075119

Spriggs EL, Christin P-A, Edwards EJ. 2014. C4 photosynthesis promoted species diversification during the Miocene grassland expansion. *PLoS One* 9:e97722.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.

Stebbins GL. 1985. Polyploidy, hybridization, and the invasion of new habitats. *Ann. Mo. Bot. Gard.* 72:824.

Stergachis AB, Neph S, Sandstrom R, Haugen E, Reynolds AP, Zhang M, Byron R, Canfield T, Stelhing-Sun S, Lee K, et al. 2014. Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature* 515:365–370.

Stiehler F, Steinborn M, Scholz S, Dey D, Weber APM, Denton AK. 2021. Helixer: cross-species gene annotation of large eukaryotic genomes using deep learning. *Bioinformatics* 36:5291–5298.

Stitzer MC, Seetharam AS, Scheben A, Hsu S-K, Schulz AJ, AuBuchon-Elder T, El-Walid M, Ferebee TH, Hale CO, La T, et al. 2025. Extensive genome evolution distinguishes maize within a stable tribe of grasses. *bioRxiv* [Internet]:2025.01.22.633974. Available from: https://www.biorxiv.org/content/10.1101/2025.01.22.633974v1.abstract

Tu X, Mejía-Guerra MK, Valdes Franco JA, Tzeng D, Chu P-Y, Shen W, Wei Y, Dai X, Li P,

969 Buckler ES, et al. 2020. Reconstructing the maize leaf regulatory network using ChIP-seq
970 data of 104 transcription factors. *Nat. Commun.* 11:5089.

971 Tu X, Ren S, Shen W, Li J, Li Y, Li C, Li Y, Zong Z, Xie W, Grierson D, et al. 2022. Limited
972 conservation in cross-species comparison of GLK transcription factor binding suggested
973 wide-spread cistrome divergence. *Nat. Commun.* 13:7632.

974 Voichek Y, Hristova G, Mollá-Morales A, Weigel D, Nordborg M. 2024. Widespread
975 position-dependent transcriptional regulatory sequences in plants. *Nature Genetics*:1–9.

976 Wasserman WW, Sandelin A. 2004. Applied bioinformatics for the identification of regulatory
977 elements. *Nat. Rev. Genet.* 5:276–287.

978 Weirauch MT, Hughes TR. 2011. A Catalogue of Eukaryotic Transcription Factor Types, Their
979 Evolutionary Origin, and Species Distribution. In: Hughes TR, editor. A Handbook of
980 Transcription Factors. Dordrecht: Springer Netherlands. p. 25–73.

981 Wittkopp PJ, Haerum BK, Clark AG. 2004. Evolutionary changes in cis and trans gene
982 regulation. *Nature* 430:85–88.

983 Woodhouse MR, Cannon EK, Portwood JL 2nd, Harper LC, Gardiner JM, Schaeffer ML, Andorf
984 CM. 2021. A pan-genomic approach to genome databases using maize as a model system.
985 *BMC Plant Biol.* 21:385.

986 Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.*
987 8:206–216.

988 Xu H, Zhang Z, Zhao Q, Gao Y, Xiang Y, Chai J, Li Y, Hou X. 2024. Study on molecular
989 response of alfalfa to low temperature stress based on transcriptomic analysis. *BMC Plant*
990 *Biol.* 24:1244.

991 Yamasaki K, Kigawa T, Inoue M, Watanabe S, Tateno M, Seki M, Shinozaki K, Yokoyama S.
992 2008. Structures and evolutionary origins of plant-specific transcription factor
993 DNA-binding domains. *Plant Physiol. Biochem.* 46:394–401.

994 Yang S, Oksenberg N, Takayama S, Heo S-J, Poliakov A, Ahituv N, Dubchak I, Boffelli D.
995 2015. Functionally conserved enhancers with divergent sequences in distant vertebrates.
996 *BMC Genomics* 16:882.

997 Zeng R, Shi Y, Guo L, Fu D, Li M, Zhang X, Li Z, Zhuang J, Yang X, Zuo J, et al. 2025. A
998 natural variant of COOL1 gene enhances cold tolerance for high-latitude adaptation in
999 maize. *Cell* [Internet] 0. Available from:
1000 http://www.cell.com/article/S0092867424014314/abstract

1001 Zhang C, Mirarab S. 2022. ASTRAL-Pro 2: ultrafast species tree reconstruction from multi-copy
1002 gene family trees. *Bioinformatics* 38:4949–4950.

1003 Zhang T, Huang W, Zhang L, Li D-Z, Qi J, Ma H. 2024. Phylogenomic profiles of

whole-genome duplications in Poaceae and landscape of differential duplicate retention and losses among major Poaceae lineages. *Nat. Commun.* 15:3305.

Zhao H, Sun Z, Wang J, Huang H, Kocher J-P, Wang L. 2014. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* 30:1006–1007.

Zhao L, Xie L, Zhang Q, Ouyang W, Deng L, Guan P, Ma M, Li Y, Zhang Y, Xiao Q, et al. 2020. Integrative analysis of reference epigenomes in 20 rice varieties. *Nat. Commun.* 11:2658.