A map of integrated cis-regulatory elements enhances gene regulatory analysis in maize

Nicolás Manosalva Pérez[1,2,*], Jasper Staut[1,2,*], Andrés Matres Ferrando[1,2], Indeewari Dissanayake[1,2] and Klaas Vandepoele[1,2,3,+]

(1) Ghent University, Department of Plant Biotechnology and Bioinformatics, Technologiepark 71, 9052 Ghent, Belgium
(2) VIB-UGent Center for Plant Systems Biology, Technologiepark 71, 9052 Ghent, Belgium
(3) VIB Center for AI & Computational Biology, VIB, Ghent, Belgium
(*) These authors contributed equally to this work
(+) Corresponding author; Klaas Vandepoele (klaas.vandepoele@psb.vib-ugent.be)

Running Title

A map of integrated CREs enhances GRN analysis in maize

Abstract

Cis-regulatory elements (CREs) are non-coding DNA sequences that modulate gene expression. Their identification is critical to study the transcriptional regulation of genes controlling key traits that govern plant growth and development. They are also crucial components for the delineation of gene regulatory networks, which represent regulatory interactions between transcription factors (TFs) and target genes. In maize, CREs have been profiled using different computational and experimental methods, but the extent to which these methods complement each other in identifying functional CREs is unclear. Here, we report the data-driven integration of different maize CRE profiling methods to optimize the capture of experimentally-confirmed TF binding sites, resulting in maps of integrated CREs (iCREs) showing increased levels of completeness and precision. We combined the iCREs with a wide diversity of gene expression data under drought conditions to perform motif enrichment and infer drought-specific GRNs. Mining these organ-specific GRNs revealed previously characterized and novel candidate regulators of maize drought response. Furthermore, by studying the transposable elements (TEs) overlapping with iCREs, we identified few TE superfamilies displaying typical epigenetic features of regulatory DNA that are potentially involved in wiring specific TF-target gene regulatory interactions. Overall, our study showcases the integration of different omics data sources to generate a high-quality collection of CREs, together with their applicability to better characterize gene regulation in the complex maize genome.

Introduction

The characterization of functional genomic elements, particularly cis-regulatory elements (CREs), is vital to understand the regulation of complex gene expression programs that coordinate the development and environmental adaptation in plants (Marand et al. 2023). CREs contain short and specific non-coding DNA sequences that are bound by transcription factors (TFs) to regulate the expression of genes. These sequences are also termed TF binding sites (TFBSs) or motifs. The comprehensive identification of CREs is key to study gene regulation and identify gene regulatory networks (GRNs), which represent regulatory interactions between TFs and their target genes. GRNs have been widely used to discover the molecular regulators in signaling pathways that regulate plant growth and response to stresses (Sullivan et al. 2014; Gaudinier et al. 2018; Maher et al. 2018; D. Chen et al. 2018; Reynoso et al. 2019; Zhou et al. 2020; Zander et al. 2020; Kajala et al. 2021; De Clercq et al. 2021; Clark et al. 2021; Depuydt, De Rybel, and Vandepoele 2023). In maize, the second most-produced crop worldwide, the characterization of CREs and inference of GRNs is challenging given its large and repeat-rich genome. The maize genome has long intergenic regions, long-range and distal regulatory interactions, and consists of 80% transposable elements (TEs) (Jiao et al. 2017; E. Li et al. 2019; Z. Lu et al. 2019; Ricci et al. 2019). TE proliferation during maize evolution shaped its genome function and structure, and although they are mostly silenced and inactive, they can influence gene expression if they carry CREs upon their transposition in the proximity of genes (Marand et al. 2023). Additionally, the insertion of TEs can also disrupt existing CREs or alter the surrounding epigenetic context due to the spreading of repressive chromatin marks (Hirsch and Springer 2017; Schmitz, Grotewold, and Stam 2022; Marand et al. 2023).

Both experimental and computational methods can be employed to identify CREs in a genome-wide manner. Chromatin immunoprecipitation sequencing (ChIP-seq) profiles the *in vivo* binding of one specific TF in a specific context, making it a low-throughput method. The characterization of accessible chromatin regions (ACRs), a TF-independent method to identify putative CREs, can be achieved using various techniques such as assay for transposase accessible chromatin sequencing (ATAC-seq) (Buenrostro et al. 2015), DNase I hypersensitive sites sequencing (DNase-seq) (Boyle et al. 2008), micrococcal nuclease sequencing (MNase-seq) (Yuan et al. 2005), and MNase-defined cistrome-occupancy analysis (MOA-seq) (Savadel et al. 2021). These methods aid the identification

of CREs because regions depleted of nucleosomes allow TFs to bind to their cognate CREs, hinting at functional regulatory regions in a specific *in vivo* context (Noll 1974; Oudet, Gross-Bellard, and Chambon 1975; Weintraub and Groudine 1976; X. Wang et al. 2011). DNA methylation is also involved in gene regulation: hypomethylated regions tend to overlap with ACRs (Eli et al. 2016; Oka et al. 2017; Ricci et al. 2019), while hypermethylation is usually observed in silenced genomic regions (Soppe et al. 2002). Unmethylated regions (UMRs), profiled with deep whole-genome bisulfite sequencing, were shown to be located nearby genes expressed and to overlap with ACRs in tissues different from the one profiled. Furthermore, DNA methylation patterns in plants show low variability during the vegetative development (Eichten et al. 2013; Kawakatsu et al. 2016) and in response to environmental changes (Eichten and Springer 2015; Ganguly et al. 2017; Crisp et al. 2017). Thus, UMRs show potential to identify CREs that are functional in different organs/tissues, developmental stages, or environments, allowing to profile a more complete set of regulatory elements compared to ACRs (Crisp et al. 2020). Computational methods that characterize putative CREs focus mainly on profiling genome conservation to identify conserved non-coding sequences (CNSs). CREs tend to show increased sequence conservation as a result of purifying selection due to their role in gene regulation (Kimura 1991; Cooper and Brown 2008; Haudry et al. 2013). FunTFBS (Tian et al. 2020), *msa_pipeline* (Wu et al. 2022), BLSSpeller (De Witte et al. 2015), or the method developed by (Song et al. 2021) are examples of approaches that have been employed to identify CNSs in plants, including maize.

In contrast to TF ChIP-seq profiling, profiling chromatin accessibility, DNA methylation or genome conservation provides a more general, context-independent way to generate a complete map of CREs. While accurate CRE identification aids the characterization of regulatory DNA flanking individual genes, it also enables the development of motif-based tools that predict GRNs, supporting gene regulation analysis (Wilkins et al. 2016; Kulkarni et al. 2018; P. Ding et al. 2021; Ferrari, Manosalva Pérez, and Vandepoele 2022; Manosalva Pérez et al. 2024).

Despite the availability of diverse methods that characterize putative CREs in plants, it is currently unclear which one is most suited to identify TFBSs and regulatory interactions. Here we present an evaluation of four CNS detection methods in maize and their comparison and integration with UMRs and a compendium of ACRs. We benchmarked these three CRE sources to assess their complementarity and generate a map of integrated CREs (iCREs). We demonstrated the power of the iCREs in identifying

functional TFBSs by inferring iCREs-based GRNs that are specific for maize drought response across different organs. These GRNs allowed identifying diverse TFs playing a role in maize drought response. These drought GRNs also shed light on the downstream processes they control, as well as their organ specificity. Additionally, we found specific TE superfamilies that are enriched in iCREs, display chromatin signatures of regulatory DNA, are overrepresented for specific TFBSs, and influence gene expression.

Results

Complementarity between CREs identified using computational and experimental approaches

Different CRE-profiling methods exist, but an evaluation of their complementarity at predicting functional TFBSs is missing in plants. We evaluated four CNS detection methods that have been previously applied in plants (BLSSpeller, *msa_pipeline*, funTFBS and a method developed by (Song et al. 2021)), along with two experimental methods used to profile epigenetic landscapes (ACRs and UMRs), as candidate approaches to detect putative CREs. In order to evaluate and compare their ability to detect regulatory DNA, in particular TFBSs, we used a ChIP-seq gold standard that comprises binding events of 104 TFs in maize mesophyll cells (Tu et al. 2020). As chromatin accessibility indicates regions that can be bound by TFs, we considered the ACR and UMR datasets included in our benchmark as two alternative, silver standards to help us assess the robustness of our findings. To obtain a compendium of ACRs, we merged 13 high-quality datasets across different tissues (leaf, root, ear, tassel, husk, inner stem, above ground seedling, and axillary bud; see Methods; Figure S1). As a control, we included repetitive DNA and TEs in our benchmark, which are known to be depleted of CREs (Marand et al. 2023).

We found that predictions of all the tested methods were significantly enriched (p-value < 0.05) for the putative binding sites of the ChIP-seq data, while the repeats/TEs were depleted (Table S1), demonstrating the ability of each method to identify TF binding events. Next, we computed the precision and recall of all methods considering the ChIP-seq dataset as ground truth. Precision is a metric for the correctness of the predictions, while recall reports the completeness (i.e., how many regions in the gold standard were

identified). We observed notable differences in performance between methods, in particular among CNS detection methods (Figure 1A). While the precision is similar between different CNS detection methods (between 9.0% and 13.3%), funTFBS shows the lowest recall (0.6%), followed by BLSSpeller, msa_pipeline (18.2% and 20.1%), and the method proposed by (Song et al. 2021) (62.9%). The method by (Song et al. 2021) was the best performing CNS detection method, both in precision and recall. To assess the robustness of these trends, we used the compendium of ACR datasets and the UMR dataset as alternative silver standards and found that (Song et al. 2021) is the best CNS detection method also for predicting ACRs and UMRs (Figure S2). However, ACRs and UMRs show better precision and recall at predicting the ChIP-seq-confirmed regions than the best-performing CNS detection method (Figure 1A). Among these two, UMRs are the best at predicting the TF binding events of the gold standard (1.7% higher precision 15.9% higher recall than ACRs). The evaluation with the ACR silver standard (where ACRs are excluded from the evaluation) further confirms that UMRs outperform all CNS detection methods (Figure S2A).

Based on our results indicating that both the CNS detection and experimental datasets identify functional regulatory DNA, we next assessed their complementarity, starting with the CNS datasets. We subdivided the CNS sets of the four methods into mutually exclusive sets of overlapping and unique genomic regions. Then, we calculated the precision, recall, and F1 of each set (Figure S3). As expected, the intersection of all four CNS detection methods has the highest precision. Next, we evaluated if the integration of different CNS detection methods can outperform the best performing individual method. Therefore, we took the CNS subset with the highest precision and subsequently added the set with the next best precision until all CNS detection methods were included. Evaluating precision, recall and F1 (the harmonic mean between precision and recall) at each stage ("cumulative" F1) shows that an ensemble approach yields a set of CNSs with a much higher precision than the best individual CNS set, albeit at the cost of a reduced recall (Figure 1B; Figure S3). However, when comparing the CNSs of (Song et al. 2021) with an ensemble of similar recall, both have similar performances (the ensemble's precision is 0.0006% higher and its recall is 0.008% higher). We observed that the best CNS ensemble only shows a marginal improvement in F1 (0.001% higher) over the best individual method, demonstrating that the ensemble is not able to outperform the method of (Song et al. 2021). Thus, for simplicity, we did not continue with a CNS ensemble approach. Instead, we used the CNSs of (Song et al. 2021) to assess the complementarity

between sequence conservation (CNSs), chromatin accessibility (ACRs) and DNA methylation (UMRs) for the purpose of integrating them into a map of putative CREs.

To integrate CNSs, ACRs and UMRs in a way that maximizes the agreement with the ChIP-seq gold standard and thereby optimally captures functional CREs, we performed a complementarity analysis, similar as with the CNS detection methods. Mutually exclusive sets of genomic regions were defined by overlapping the different methods and calculating the performance for each of them (Figure 1C). Each of the three methods contains a substantial fraction of unique regions. Regions unique to UMRs show a better agreement with ChIP-seq-confirmed bindings sites than unique ACRs or CNSs, as well as regions where ACRs and CNSs overlap but do not coincide with a UMR. In addition, we observe that combining methylation data with other methods (CNSs or ACRs) can further improve precision (up to 37%, compared to 20% when considering all UMRs; Figure 1C). This result reveals that genome conservation, chromatin accessibility and DNA methylation contain complementary information to predict CREs. Finally, we integrated the CNSs, ACRs and UMRs into a set of “integrated CREs” or iCREs. As described before, we combined the mutually exclusive overlapping and unique sets into ensembles, and selected the one with the highest “cumulative” F1 score (Figure 1C). This ensemble, referred to as the “max F1 iCREs”, is equivalent to the intersection between the UMRs and the ACRs. We then constructed a second ensemble consisting of the union of CNSs, ACRs and UMRs that contains the regions of all three methods, which will be referred to as the “all iCREs”. Figure 1D indicates the distribution of the ensemble sets in different genomic regions and/or features. As expected, the “max F1 iCREs” have lower genome coverage (48 Mb) and are more often located in a UTR or proximal region (19% UTR and 28% proximal), while “all iCREs” cover more of the genome (184 Mb) and have a lower proportion in these regions (9% UTR and 20% proximal). In summary, we show that DNA methylation has a high level of agreement with ChIP-seq-confirmed TFBSs and that it can be complemented with chromatin accessibility and/or genome conservation information. This complementarity allows us to generate a high confidence (max F1 iCREs) and a more comprehensive (all iCREs) map of putative maize CREs.

Inference of drought-responsive gene regulatory networks using the iCREs and a organ-specific gene expression atlas

iCREs represent a comprehensive collection of maize regulatory regions, having the potential to improve GRN inference. Therefore, we implemented a framework to perform iCRES-based motif enrichment and GRN inference starting from a set of functionally-related genes, using the MINI-AC tool (Manosalva Pérez et al. 2024) (Figure 2A).

The frequency and severity of droughts is expected to increase given the current climate emergency, ultimately impacting the yield of maize. To find candidate drought regulators and understand the processes they control, we first constructed a gene expression atlas by selecting and curating 20 maize drought RNA-seq studies to obtain sets of drought-responsive differentially expressed genes (DEGs; Table S2, Figure S4, Materials and Methods). These studies include samples of six different organs: leaf, root, seminal root, primary root, kernel, and ear. Next, we identified “marker” genes that respond to drought exclusively in each of these organs, as well as general or “core” drought-responsive genes that consistently respond to drought across various experimental conditions, samples, and genotypes.

The “core” drought-responsive genes were identified by counting the number of contrasts (the comparison between a control and drought sample) in which each drought-responsive gene is detected as differentially expressed (DE). Next, we selected the genes up- or down-regulated in at least one third of the maximum number of contrasts that share at least one DEG. This yielded 568 “core UP” genes in 15 contrasts and 1702 “core DOWN” genes in 10 contrasts (Figure 2B). Given the diverse selection of studies, there were genes showing repression (i.e. down-regulation) in some contrasts and activation (i.e. up-regulation) in others. Therefore, we quantified the activation or repression consistency of the “core” genes (see Methods; consistency score of 0.5 = fluctuating expression, consistency score of 1 = consistent expression). While we overall observed high consistency scores, they were higher for the down-regulated core genes (median 0.88) than for the up-regulated ones (median 0.75). This trend likely stems from the leaf samples, as this is the organ with most contrasts and the only one replicating low activation consistency scores (Figure 2B and S5). This difference in consistency scores might be due to the drought-induced repression of numerous photosynthetic genes, resulting in higher repression consistency scores (Berrío et al. 2022).

The organ-specific drought-responsive genes were defined as the DEGs in more than half of the contrasts of that organ (Figure 2C; kernel was discarded due to the large number of DEGs and having only one contrast). We evaluated the overlap between these organ-specific activated and repressed genes to define unique or “marker” drought-responsive genes. We found that in most organs, the majority of the specific DEGs are unique (Figure 2C). For instance, leaf and root have 85% and 87% of unique repressed DEGs, respectively. Clustering the “core” drought-responsive genes based on their expression log fold-change (logFC) confirmed a ubiquitous drought response (repression and activation across many contrasts). Conversely, clustering of the “marker” drought-responsive genes supported their specificity (repression/activation in corresponding organ contrasts; Figure 2D).

Next, for each drought-responsive gene set, we identified their associated iCREs and performed motif enrichment and GRN inference using MINI-AC. The construction of the GRN is done by finding motifs overrepresented in the iCREs associated to a gene set and linking them to the TFs that bind those motifs (MINI-AC normally does this with a set of ACRs). We evaluated if there was an advantage in performing motif enrichment using iCREs as input (iCREs-based enrichment) compared to using the adjacent non-coding regions and introns of the drought-responsive genes (default promoter-based enrichment, see Materials and Methods). Using the “all iCREs” set reduces the number of genome-wide motif occurrences of the MINI-AC TFBS collection by 90%, potentially reducing false positives. Therefore, we compared the performance of these two approaches by measuring how well MINI-AC’s enrichment analysis predicts the motifs bound by a gold standard set of TFs, which are the DE TFs in each drought-responsive gene set. In other words, we evaluate if the motifs associated to iCREs of drought-responsive DEGs are significantly enriched for drought-responsive TFs. Additionally, we also made a comparison of the “all iCREs” and “max F1 iCREs” sets (Figure 3A). Our results showed that the iCREs-based motif enrichment outperforms the promoter-based enrichment in 9 out of the 12 gene sets for the two metrics tested (F1 and the area under the precision recall curve or AUPRC, that measures the performance of the top-scoring predictions; See Materials and Methods). The iCREs-based approach achieves a median F1 10% higher and a median AUPRC 6% higher than the promoter-based approach. Although the “maxF1” and “all” iCREs showed similar performances, the latter was superior for most of the drought-responsive gene sets, with a median 0.049 AUPRC and 0.073 F1, and a median 0.065 AUPRC and 0.1 F1, respectively. The inferred GRN of the “core” genes

predicted with the promoter-based method had a similar number of unique TFs as the GRN of the iCREs-based approach (1194 and 1160 respectively), but four times more edges (5.44 million and 1.37 million respectively) and two times more unique target genes (6440 and 3106 respectively). Given that the iCREs-based approach reduces the genomic space used for finding TFBS to regions with experimental or computational evidence of being functional, this decrease in edges and target genes most likely implies a reduction of false positive predictions.

Next, we analyzed the drought-responsive GRNs using “all iCREs” and the “marker” gene sets, representing organ-specific drought DEGs. Using MINI-AC, motifs were scored based on how overrepresented they are in the iCREs of the DE genes to compute a motif enrichment rank. Comparing the ranks along with the activation and repression consistency in each organ allowed us to pinpoint specific drought stress regulators (Figure 3B). We observed high expression and motif enrichment specificity for WRKY TFs in root cell types, especially in seminal root (WKRY57, WRKY118, WKRY108, and WKRY49). Several WRKY TFs have been characterized for their role in maize drought response (L. Zhao et al. 2021; C.-T. Wang et al. 2018) and have been associated with an increase of lateral roots (Gulzar et al. 2021). In leaf, we observed high expression specificity of Trihelix TFs (Trihelix19 and 12) as well as a low (but not specific) motif enrichment rank. The Arabidopsis ortholog of these TFs, GT1, has been characterized for its role in the repression of cell growth (Caro, Desvoyes, and Gutierrez 2012; Breuer et al. 2012) and the regulation of drought tolerance (Yoo et al. 2010), which is in line with the growth arrest that occurs under drought stress (Dubois and Inzé 2020). bZIP95 in leaf also showed both high expression and motif enrichment specificity, hinting at an important role in the drought response of leaf (Figure 3B). Focusing on leaf, we explored the biological processes that the predicted DE regulators are controlling. We found an expected repression of photosynthesis- and carbon metabolism-related genes under the control of repressed TFs. Among these, we find *Arabidopsis thaliana* orthologs (PIL6, MYBH or CRF5) that have been characterized in light signaling and cell expansion (Fujimori et al. 2004; D. Lu et al. 2014; Raines et al. 2016). Conversely, we found that the aforementioned Trihelix19, Trihelix12 and bZIP95 are controlling genes involved in response to abscisic acid and water deprivation, further supporting their potential role in maize leaf drought response (Figure 3C). Taken together, these results illustrate the potential of employing the iCREs to infer context-specific GRNs and identify candidate regulators.

iCREs-guided characterization of the regulatory role of different types of transposable elements

While numerous examples of TEs influencing gene expression have been reported in plants (Hirsch and Springer 2017; Noshay et al. 2021; Deneweth, Van de Peer, and Vermeirssen 2022), the role of specific types of TEs contributing to regulatory elements is less clear. Therefore, we identified a subset of TEs fully contained within iCREs and examined their potential regulatory role based on the superfamily they belong to. In the maize genome, different TE superfamilies and types of repetitive elements are classified according to their sequence similarity and transposition specificity (Wicker et al. 2007) (see Materials and Methods). We assessed whether specific TE superfamilies/repetitive elements were significantly overrepresented in the iCRES. While the full set of TEs is depleted in the iCREs, the superfamilies Tc1/Mariner, PIF/Harbinger, and hAT that belong to the TE order of terminal inverted repeats (TIR), show a high enrichment in iCREs (enrichment folds of 3.5, 2.9, and 2.3, respectively; Figure 4A). Additionally, the helitron superfamily and two superfamilies belonging to the long interspersed nuclear element (LINE) TE order also showed a 1.4- to 2.2-fold enrichment. Using the non-coding and non-TE genomic space as control regions confirms that the iCREs are depleted from TEs, except for these 6 superfamilies. Additionally, a gene ontology (GO) enrichment analysis revealed that the TEs within iCREs belonging to the TIR superfamily are the most significantly associated with specific groups of functionally-related genes (Figure 4B). For example, we found that the responses to (a)biotic stresses such as salt, cold, bacteria, and fungus are among the most enriched GO terms (q-values between 5.5×10^{-3} and 2.4×10^{-8} and enrichment folds between 1.3 and 1.4), along with developmental functions like regulation of flower development and chloroplast organization. While other TE superfamilies of the LTR and LINE orders also yielded GO enrichment among their associated genes (mostly related with carbohydrate metabolism; Table S3), Figure 4B depicts only the terms with most significant enrichment (q-value < 0.008 with >100 genes in that category).

Next, to investigate if the epigenetic profile of the TEs within iCREs is different from the one of the TEs outside iCREs, we compared their overlap with different chromatin states (CS). CSs are sets of genomic regions with specific combinations of epigenetic features that include histone marks, histone variants, DNA methylation, ACRs, and binding sites of

chromatin-associated factors (Liu et al. 2018). We observed that TEs within iCREs share a similar CS enrichment pattern as the iCREs, which implies that they tend to be in promoter, intronic, UTR, and intergenic regions (Figure 4C). They also display the opposite pattern as the full set of TEs and the TEs outside iCREs, that are enriched in CSs associated with repetitive regions showing DNA methylation. Furthermore, we explored the CS enrichment pattern per TE superfamily and repetitive element type, in, and outside iCREs. We found that the helitron and TIR TEs within iCREs, in particular Tc1/Mariner, CACTA, PIF/Harbinger, Mutator, and hAT have the most similar CS enrichment pattern to the iCREs (Figure S6), suggesting they represent functional regulatory DNA.

Besides examining the epigenetic profile of the TEs within iCREs, we also studied expression levels of nearby genes. We first evaluated whether the distribution of the distance to the closest gene was different for the TEs that overlap with iCREs compared to those that do not. We found that the proportion of TEs that are closer to genes (between 50 and 2000 base pairs (bps)) is higher when they are within iCREs than when they are outside (Figure S7). Considering this, we used the gene expression atlas of maize drought to evaluate differences in the expression of genes associated with TEs within iCREs, outside iCREs, or genes associated with both. This analysis revealed that in both control and drought samples, the median expression of genes associated with TEs in iCREs is significantly higher than those outside iCREs at distances of less than 2000 bps from the genes or when they are inside the gene body, meaning TEs within introns and UTRs (median TPMs 7.5-8.1 in iCREs vs. 4.8-5.2 outside iCREs at distance 0 from genes; p-value<0.0001 using a one-sided Mann-Whitney U test; Figure 4D; Table S4). This trend is also noticeable for the expression values in the control and drought samples when only considering the genes associated with the Tc1/Mariner and PIF/Harbinger superfamilies that display enrichment in CSs typical of regulatory DNA (Table S4).

Finally, we employed MINI-AC to perform motif enrichment on the PIF/Harbinger and Tc1/Mariner TEs fully overlapping with iCREs, and whose associated genes were functionally annotated with “response to salt stress”, “defense response to bacterium”, or “regulation of flower development” (most enriched GO categories; Figure 4E). We found that the MYB TFs and the high mobility group (HGM) TFs were enriched in all the different sets of TEs within iCREs we tested. The MYBs were especially enriched in the PIF/Harbinger TEs associated with genes involved in flower development, and the HGM in the Tc1/Mariner TEs associated with genes involved in response to salt stress. Additionally, in the latter, the SBP family also showed high enrichment. The bHLH TF

family was highly enriched in the PIF/Harbinger TEs in the context of flower development. In the Tc1/Mariner TEs associated with genes involved in response to bacteria, there was a high enrichment of WRKY TFs, which is expected for this family, although this was not the case for the PIF/Harbinger TEs (X. Chen et al. 2019). Next, we investigated whether *in vivo*-profiled binding sites of the aforementioned TF families were overrepresented within the PIF Harbinger and Tc1-Mariner TE superfamilies. We achieved this by performing enrichment analysis on TF ChIP-seq peaks from (Tu et al. 2020). For example, we found that the MYB family had an enrichment π-value (see Materials and Methods) of 35.5 in the PIF Harbinger TEs in iCREs, compared to 6.9 outside iCREs. A similar trend was also found for the TF families bHLH (31.5 inside iCREs vs. 6.5 outside). For the Tc1 Mariner superfamily, we found that the MYB and WRKY families had a π-value of 22.9 and 19.8 in iCREs, compared to 4 and 3.9 outside iCREs, respectively. This observation reveals that specific TE superfamilies part of iCREs harbor binding sites bound by specific TFs.

Altogether, these results highlight the potential regulatory role of specific TE superfamilies in maize. The presence of overrepresented TFBSs in these TE superfamilies, as well as the prevalence of specific biological processes in the associated genes, suggests that the wiring of these TF families, linking TFs to their target genes, is (partly or completely) mediated by specific TEs.

Discussion

Integrating datasets from various regulatory DNA profiling methods in plants holds significant potential for enhancing our understanding of gene regulation. CRE-profiling methods such as ChIP-seq can detect direct TF binding, but are relatively low-throughput. Profiling chromatin accessibility, DNA methylation or genome conservation offers a more straightforward way of generating a complete map of CREs. Although previous studies have focused on the agreement between ChIP-seq, CNSs, ACRs and UMRs, their complementarity and potential for integration was largely unexplored (Ricci et al. 2019; Crisp et al. 2020; Song et al. 2021; Marand et al. 2023). We show that regions supported by multiple methods are more likely to be confirmed by ChIP-seq, and that each method contains a substantial amount of unique regions, indicating complementarity. While genome conservation provides increased confidence in a region's regulatory role, our

benchmark demonstrates that it is less reliable than profiling methylation or accessibility. Indeed, CNSs are prone to false positives, such as conserved non-coding regulatory RNA, and false negatives, e.g., when a TFBS has been recently introduced in evolutionary time, thereby lacking conservation, or TFBSs in TEs since these regions are usually masked when determining sequence conservation (Siepel et al. 2005; Haudry et al. 2013; Van De Velde, Heyndrickx, and Vandepoele 2014). Among the included CNS methods, we show that the CNS method by (Song et al. 2021) provides the best results. By covering a smaller evolutionary distance than other methods (only species within the *Andropogoneae* tribe), it might be able to detect more recent TFBSs, enhancing recall, while maintaining a good precision. Following the CNS methods, we find chromatin accessibility as the next best approach to identify functional CREs. However, ACRs might miss the TFBSs of pioneering TFs, which don't need accessibility to perform their function (Lai et al. 2021; Strader, Weijers, and Wagner 2022).

Organ- or tissue-specificity in our datasets should be taken into account when interpreting our results. Our benchmark puts forward UMRs as the most informative for predicting TFBSs. However, a reason for the superior performance of UMRs might be that they were profiled in maize leaf, similar to our benchmark's ChIP-seq gold standard. While UMRs are considered relatively stable across organs/tissues, some degree of specificity remains, leading to a potential overestimation of their performance (Crisp et al. 2020). The ACRs we compiled combine accessibility data from different organs and tissues, and are therefore not expected to be strongly specific. Notably, a single-cell ATAC-seq dataset was included (Marand et al. 2021), representing a high-resolution set of ACRs of different organs and cell types. However, as not every cell type, developmental stage, and condition is included in our ACR compendium, this dataset is expected to still exhibit some context-specific biases (e.g., missing of CREs associated with specific *in vivo* contexts). In contrast, genome conservation is fully context-independent. The leaf-specificity of the gold standard ChIP-seq experiments could result in an underestimation of CNS performance when considering all CREs in the maize genome. Indeed, we do find some evidence for this hypothesis in the additional benchmark against the less organ-specific UMR silver standard, where CNSs are outperforming ACRs. This indicates that CNSs might not necessarily be less informative than ACRs when evaluated in a more context-independent manner.

Assembling the iCREs raised the opportunity to extend MINI-AC's functionality to infer context-specific GRNs using the iCREs and a list of genes. Currently, the state-of-the-art

methods to infer GRNs have a strong focus on exploiting single-cell (multi-)omics datasets because they offer an unprecedented level of resolution (Jiang et al. 2022; Ferrari, Manosalva Pérez, and Vandepoele 2022; Kamimoto et al. 2023; Bravo González-Blas et al. 2023; Fleck et al. 2023). However, those protocols are still not optimized for every model species and/or experimental condition, and they are also expensive, which might make them suboptimal or unnecessary, depending on the biological question being addressed. Conversely, the inference of GRNs from a specific list of genes can aid in highlighting candidate regulators in a simple and straightforward manner. Such tools exist, like TF2Network, TDTHub, ConnecTF, or PlantRegMap (Kulkarni et al. 2018; Tian et al. 2020; Brooks et al. 2020; Grau and Franco-Zorrilla 2022). However, they test motif enrichment by applying a hypergeometric or Fisher's exact test to TFBSs in pre-defined genomic windows adjacent to genes (e.g., upstream promoter region), which misses distal CREs. In our iCREs-based framework, that employs MINI-AC, the motif enrichment encompasses the whole non-coding genome, but reduces potential false positive TFBSs. We showed that this increases the performance in the prediction of expected regulators. We found that the "maxF1" iCREs set, which is the overlap between the UMRs and the ACRs, has the best agreement (F1 metric) with the leaf ChIP-seq gold standard. The "all" iCREs set, containing all UMRs, ACRs and CNSs, has a poorer agreement with the gold standard. However, when used to infer GRNs, it marginally improves predictions of enriched binding sites for the DE regulators in an input set of genes. This, however, is not completely unexpected, given that the gold standard is derived from mesophyll cells in normal conditions, and we predict drought GRNs for multiple organs. Thus, the aforementioned elements unique to the "all" iCREs set might contain additional relevant CREs not captured by the ChIP-seq mesophyll gold standard.

The regulatory role of PIF/Harbinger TEs found in our analysis has been previously explored and confirmed in *Brassica oleracea*. In this species, insertions of TEs belonging to this superfamily altered the expression of a TF responsible for purple traits in cultivars. Additionally, reporter assays showed that promoter sequences containing this TE insertion induced higher expression than those that do not (X. Li et al. 2024). These results are in line with our findings, where we observed that genes close to both TEs within and outside iCREs have higher expression levels than those close to only TEs outside iCREs. This indicates that those TEs have a promoting effect on gene expression rather than a repressing effect. The role of TEs on stress gene regulatory networks in *Arabidopsis thaliana* and tomato has previously been assessed, where enriched motifs were found in

different TE superfamilies, similar as in our study; e.g., WRKY TFs in Mariner TEs (Deneweth, Van de Peer, and Vermeirssen 2022). In addition, the contribution of TEs to the stress response in maize has been studied before (Le et al. 2014; Makarevitch et al. 2015; Roquis et al. 2021), as well as the potential regulatory role of maize TEs within ACRs (H. Zhao et al. 2018; Noshay et al. 2021). Our findings provide support for previously proposed models where TEs play a role in rewiring GRNs (Britten and Davidson 1971; Feschotte 2008; Rebollo, Romanish, and Mager 2012). While this phenomenon has been observed in animals and some plants (T. Wang et al. 2007; Kunarso et al. 2010; Lynch et al. 2011; Batista et al. 2019), our results hypothesize that TIR TEs play a similar role in wiring maize stress response networks.

In our study, the assumption that one iCRE controls its closest gene fails to capture long-range regulatory interactions, where adjacent genes are skipped, or cases where one iCRE controls more than one gene. Eventually, the generation of high-quality Hi-C (Lieberman-Aiden et al. 2009) and ChIA-PET (Fullwood et al. 2009) datasets in maize for different conditions would allow integrating them in our framework to resolve these limitations. Nonetheless, in maize, these techniques have only been applied in few conditions, limiting their general usage for predictions in other organs/tissues, conditions, or genotypes, given that these long-range interactions can be tissue-specific (Marand et al. 2021). Moreover, the innovation of a gold standard of CREs in maize that is context-independent and contains minimum false positives could improve the data-driven optimization and integration of iCREs. On top of that, iCREs hold the potential to be integrated with data of natural genetic variation associated with phenotypic changes, to explore the role of non-coding variants within TFBSs and GRNs in determining complex traits.

In summary, our study highlights the potential of integrating data derived from different methods to aid regulatory genome analysis in different biological contexts . By combining epigenetics and genomics data, we generated a comprehensive collection of maize CREs, supporting the scientific community working in maize, and highlighting the complexity and intricacies of regulatory DNA in complex plant genomes.

Materials and Methods

CNS inference

The genomic coordinates of the CNSs from (Song et al. 2021) were downloaded from https://genome.cshlp.org/content/suppl/2021/06/22/gr.266528.120.DC1. The pairwise CNS coordinates files were concatenated and merged (*bedtools merge;* version 2.2.28) (Quinlan and Hall 2010). Coordinates were then converted from version 4 to version 5 of the maize B73 reference genome using the UCSC liftOver tool (Hinrichs et al. 2006) and a version 4 to version 5 chain file available via maizeGDB (www.maizegdb.org) (Woodhouse et al. 2021). Genome-wide maize CNSs predicted by funTFBS were downloaded via the Plantregmap portal (http://plantregmap.gao-lab.org) (Tian et al. 2020) and converted from version 3 to version 4 using the online Assembly Converter tool of Ensembl plants for maize (Martin et al. 2023). Subsequently, they were converted from version 4 to version 5 using liftOver as described above.

BLSSpeller (De Witte et al. 2015) and *msa_pipeline* (Wu et al. 2022) were run using a species set covering the PACMAD clade, including species *Cenchrus purpureus*, *Oropetium thomaeum*, *Saccharum spontaneum*, *Setaria italica, Sorghum bicolor*, *Zea mays* and *Zoysia japonica* ssp. *Nagarizaki*. To obtain orthologous groups (OGs) and a species tree, OrthoFinder (Emms and Kelly 2019) version 2.5.4 was used with default settings, using proteome files consisting of longest transcripts as input (retrieved from PLAZA version 5.0) (Van Bel et al. 2022). The resulting OGs were filtered for groups that contain genes of at least two species, including maize B73. The genome FASTA files used by BLSSpeller and msa_pipeline were retrieved from PLAZA version 5.0. Repeats were masked using a k-mer-based approach described by (Song et al. 2021), using KAT (Mapleson et al. 2017). For BLSSpeller, a search space of 2 kb upstream of the translation start site (TrSS), 1 kb downstream of the translation end site (TrES) of all the genes, including the introns was used. Next, the coding sequences were masked. BLSSpeller (source code available at https://bitbucket.org/dries_decap/bls-speller-spark) assembly version 1.1 was used in alignment-free mode on the upstream, downstream and intronic regions separately, using 8 bps as motif length, using the full IUPAC alphabet of degenerate nucleotides, and a maximum of 3 degenerate characters in a motif. A confidence score cutoff of 85 was used, the minimal number of conserved OG counts was

set to 25, and the used range of BLS thresholds was [50, 60, 70, 75, 80, 85, 90, 92, 95, 97, 99]. *Msa_pipeline* was used as described by (Wu et al. 2022) (available at https://bitbucket.org/bucklerlab/msa_pipeline), using their relaxed LAST parameter set. The initial LAST step using lastdb was run with parameter -u RY32 to decrease runtime. After pairwise alignment, alignments were chained and netted (Kent et al. 2003) followed by multiple genome alignment by ROAST (https://www.bx.psu.edu/~cathy/toast-roast.tmp/README.toast-roast.html). *Zea mays* was used as the reference species. If a comparator species was polyploid (*Triticum aestivum*, *Triticum turgidum* and *Saccharum spontaneum*), each subgenome was considered as a separate species. An adaptation of the *msa_pipeline*'s conservation calling procedure was used, where conserved elements were called per chromosome using GERP++ (Davydov et al. 2010). Conservation scores were generated using *gerpcol*, and conserved elements using *gerpelem*.

Experimental regulatory datasets

The dataset containing ChIP-seq summits of 104 TFs in maize was obtained from (Tu et al. 2020), and a region of 10 bp upstream and 10 bp downstream around a summit (total of 21 bp) was considered the peak region. The peak regions of all individual TFs were concatenated and merged (*bedtools merge*). The UMRs genomic coordinates were downloaded from the supplementary dataset 3 of (Crisp et al. 2020). Publicly available ACR datasets were collected from GEO and PlantCADB (https://bioinfor.nefu.edu.cn/PlantCADB/; (K. Ding et al. 2022)). Thirteen high-quality ACR datasets were selected by omitting negative control experiments (naked DNA) and removing datasets that, similar to negative control samples, showed a high fraction (>70%; Figure S1) of distal regions (further than 2 kb from a gene), indicating low quality. Processed BED files were collected from GEO for GSE120304 (leaf and ear) (Ricci et al. 2019) and GSE155178 (the provided peak file was converted to BED) (Marand et al. 2021). The datasets GSE128434 (Z. Lu et al. 2019), GSE85203 (Z. Lu et al. 2017), GSE94291 (Oka et al. 2017), GSE97369 (Burgess et al. 2019), PRJNA382414 (H. Zhao et al. 2018), PRJNA391551 (Dong et al. 2017), PRJNA518749 (Tu et al. 2020), and PRJNA599454 (Sun et al. 2020) were downloaded from PlantCADB (K. Ding et al. 2022). Replicates per experiment were combined by intersecting them with *bedtools intersect*. All the ACR datasets were concatenated and merged (*bedtools merge*) into a single cross-tissue ACR dataset. The coordinates of all the datasets were converted from version 4 to

version 5 of the maize B73 reference genome using the UCSC liftOver tool and a version 4 to version 5 chain file available via maizeGDB (www.maizegdb.org).

Performance evaluation

The comparison of a query and a reference dataset containing genomic coordinates was done by computing, precision, recall and F1 as follows:

$$Precision = \frac{size\ of\ overlap\ between\ query\ and\ reference\ (bps)}{size\ of\ query\ (bps)}$$

$$Recall = \frac{size\ of\ overlap\ between\ query\ and\ reference\ (bps)}{size\ of\ reference\ (bps)}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

Performance was calculated, considering only the non-coding genome by first intersecting (*bedtools intersec*t) query and reference regions with the non-coding genome (GFF downloaded from PLAZA Monocots 5.0), after which the formulas above were applied.

Computation of enrichment between sets of genomic regions

The enrichment of a query set of genomics regions (for example, the CNSs) in a target set of genomic regions (for example, the ChIP-seq peaks) was tested by shuffling the latter 1,000 times across the maize non-coding or full genome. Then, the observed overlap (real overlap) between the genomic sets was compared to the overlap of the shuffled sets (overlap expected by chance). A p-value was then computed as the number of times the overlap with any of the 1,000 shuffled sets was higher than the real overlap. Enrichment fold was computed as the real overlap divided by the median of the overlap expected by chance (median of the 1,000 shuffling events).

Construction of iCRE sets

The “max F1 iCREs” set was obtained by intersecting (*bedtools intersect*) the UMRs with the cross-tissue ACR dataset (see above). The “all iCREs” set was obtained by concatenating and merging (*bedtools merge*) all UMRs, ACRs and CNSs of (Song et al. 2021). Last, the coding sequences as defined in the maize version 5 annotation and a small (59 kb) set of high confidence predicted t-RNAs (available via http://gtrnadb.ucsc.edu/), were subtracted (*bedtools subtract*) from both iCRE sets. The

genomic annotation of the regions was done with the maize GFF downloaded from PLAZA Monocots 5.0.

Processing of the gene expression atlas of maize in drought conditions

A manual curation of *Zea mays* drought RNA-seq datasets was done using CurSE (Vaneechoutte and Vandepoele 2019). In total, 20 studies, summarized in Table S2, were selected. A total of 434 samples and their corresponding FASTQ files were downloaded using *fastq-dump -I –split-files -gzip* (version 3.0.0 https://github.com/ncbi/sra-tools) and processed using the nf-core/rnaseq (Ewels et al. 2020) pipeline to process RNA-seq data. The input sample sheets were obtained using the *fastq_dir_to_samplesheet.py* script provided in the nf-core/rnaseq repository. The transcriptome, genome FASTA, and GTF files corresponding to the maize genome version 5 were downloaded from PLAZA Monocots 5.0 (Van Bel et al. 2022). Only the stages 1 (pre-processing) and 3 (pseudo-alignment and quantification) of the nf-core/rnaseq pipeline were executed using the default parameters, except for the specification of Salmon as pseudo-aligner for quantification. The index file for Salmon was obtained using *salmon index* with default parameters and without decoys (Patro et al. 2017). Samples with a mapping rate of less than 65% were discarded, as well as outlier samples within each study with a mapping rate of less than 70% (Figure S4A). Outliers in each study were samples with a mapping rate below this value: Q1 - 1.5 × (Q3-Q1) (Q1 and Q3 being the first and third quartile, respectively). The contrasting control or treatment samples of the ones discarded due to low mapping rate were also removed, leaving a total of 375 samples (141 after processing replicates).

Selection of differentially expressed genes in response to drought

Differential expression analysis was performed using *edgeR* (Robinson, McCarthy, and Smyth 2010). Genes with less than one transcript per million (TPM) in at least two samples were discarded. Next, DEGs were inferred using *glmQLFit* and *glmQLFTest*, with the multiple testing correction and the false discovery rate (FDR) calculations being done with the Benjamini-Hochberg procedure (Benjamini and Hochberg 1995). For each contrast,

the genes with an FDR < 0.05 and an absolute log-fold change (logFC) > 1 were considered DE. Of the 80 contrasts in this analysis, a large proportion did not yield any DEG, or very few (Figure S4B). Contrasts yielding less than 100 up- or down-regulated genes were considered low quality and therefore discarded, leaving a total of 55 contrasts. The activation and repression consistency scores of the DEG sets were computed as the number of contrasts in which they are up- or down-regulated, divided by the total number of contrasts in which they are DE.

iCREs-based gene regulatory networks inference

The “maxF1” and “all” iCREs sets were individually annotated to the closest gene using *bedtools closest -a $icres -b $genes -t all*. MINI-AC was employed to perform motif enrichment and GRN inference (Manosalva Pérez et al. 2024) on the iCREs associated to each drought-responsive gene set. MINI-AC was run with the genome-wide mode on the maize genome version 5, and with the “absolute base pair count” option set to “true”. An iCREs-based MINI-AC feature was added to the GitHub repository (https://github.com/VIB-PSB/MINI-AC/tree/main) to infer context-specific GRNs in maize given a list of genes. The coordinates of the neighboring non-coding regions and introns of each gene were obtained from the maize GFF file from PLAZA 5.0 and were then overlapped with the “medium” promoter file provided in the MINI-AC GitHub repository (https://github.com/VIB-PSB/MINI-AC/blob/main/data/zma_v5/zma_v5_promoter_5kbup_1kbdown_sorted.bed). To compare the promoter-based and the iCREs-based motif enrichment approaches, the precision was computed as the number of enriched motifs associated with DE TFs divided by the total number of enriched motifs. The recall was computed as the number of enriched motifs associated with DE TFs divided by the total number of motifs associated with DE TFs. F1 was computed as the harmonic mean of precision and recall. The AUPRC was computed using the function *metrics.auc* from the Python library *sklearn* (Pedregosa et al. 2011).

Transposable elements analysis and chromatin states enrichment

The coordinates of the maize transposable and repetitive elements in the genome version 5 were downloaded from MaizeGDB (https://download.maizegdb.org/Zm-B73-

REFERENCE-NAM-5.0/Zm-B73-REFERENCE-NAM-5.0.TE.gff3.gz) (Woodhouse et al. 2021). The coordinates were split into TE superfamilies based on the third column value (feature) of the GFF file. The genomic coordinates of the chromatin states were downloaded from the plant chromatin state database (http://systemsbiology.cau.edu.cn/chromstates/download.php; (Liu et al. 2018)) and converted from the AGPv3 genome version to the AGPv4 genome version, and then to the B73 RefGen v5 genome version using liftOver with default parameters (Hinrichs et al. 2006). The chain file used by liftOver was downloaded from Ensembl (Martin et al. 2023) (https://ftp.ensemblgenomes.ebi.ac.uk/pub/plants/release-55/assembly_chain/zea_mays/AGPv3_to_B73_RefGen_v4.chain.gz and https://ftp.ensemblgenomes.ebi.ac.uk/pub/plants/release-55/assembly_chain/zea_mays/B73_RefGen_v4_to_Zm-B73-REFERENCE-NAM-5.0.chain.gz). The enrichment of TEs, TEs superfamilies, and other genomic region sets in the iCREs and the chromatin states was done as described previously in this section. The shuffling was done across the non-coding genome for the iCREs enrichment and across the whole genome for the chromatin states enrichment. For visualization purposes, the enrichment metrics were integrated into the π-value, which is the −log10(p-value) × enrichment fold. The p-value was adjusted for multiple testing using the Benjamini–Hochberg method with significance level of 0.01.

The TEs fully contained within iCREs (*bedtools intersect -a $tes -b $icres -f* 1) were annotated to their closest gene (*bedtools closest -a $tes_in_icres -b $genes -t all -d*). To perform functional analysis, the genes associated with TEs were split based on the TE superfamily, and GO enrichment was performed using a gene-GO annotation file of the maize genome version 5 (https://github.com/VIB-PSB/MINI-AC/blob/main/data/zma_v5/zma_v5_go_gene_file.txt). The enrichment was calculated using a hypergeometric test and the p-value was corrected using the Benjamini-Hochberg procedure.

The iCREs linked to genes that were associated with PIF-Harbinger and Tc1-Mariner superfamilies, and that yielded enrichment in the GO terms “defense response to bacterium”, “regulation of flower development”, and “response to salt stress” were used for motif enrichment using MINI-AC with the same parameters as described earlier in this section. To evaluate the expression levels of genes associated with TEs within iCREs, the TPMs per gene and per sample of the drought gene expression atlas mentioned earlier in

this section were employed. The distance to each gene was obtained as previously mentioned using *bedtools closest*.

Author contributions

KV, NMP, and JS conceived and designed the research. JS performed the benchmark and complementarity analysis of CRE profiling methods. NMP performed the analysis of the gene regulatory networks and transposable elements. ID ran *msa_pipeline*. AMF processed the gene expression atlas of maize drought. NMP, JS, and KV wrote the manuscript with input from the rest of the authors.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to rephrase content written by the authors. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Acknowledgements

We would like to thank Dries Decap and Jan Fostier for their support and help running BLSSpeller. We also thank Chrystian Camilo Sosa Arango for developing an RNA-seq DE pipeline, used to call drought responsive DEGs. This work was supported by a Bijzonder Onderzoeksfonds grant from Ghent University (grant agreement: BOF24Y2019001901) to NMP and by The Research Foundation - Flanders (FWO; Odysseus II G0D0515N) to JS.

Supplementary datasets

Files containing the genomic coordinates of the conserved non-coding sequences (BLSSpeller, msa_pipeline, funTFBS, and Song-2021), the accessible chromatin region compendium (ACRs) and the unmethylated regions (UMRs) employed to generate the integrated cis-regulatory-elements (iCREs), and both the “all iCRE” and “maxF1 iCRE”

sets, are available at https://zenodo.org/records/13143829. These datasets are available both in version 4 and 5 of the maize genome.

Figure legends

Figure 1

Figure 1. Benchmark, complementarity assessment, and integration of CNS detection and experimental CRE-profiling methods. (A) Precision-recall plot of individual CNS detection and experimental CRE-profiling methods, using a ChIP-seq dataset as gold standard. CNS detection methods are plotted as circles, experimental CRE-profiling methods as triangles, and control datasets as squares. (B) Precision-recall plot (using ChIP-seq as gold standard) of individual CNS detection methods and different ensembles that were constructed by combining the intersections of these methods. The ensemble with the best F1 is shown in green. (C) UpSet plot of the ACRs, UMRs and CNSs, with performance metrics of the subsets. The UpSet plot shows the sizes of the full datasets on the left and the sizes of the mutually exclusive subsets on the top. Precision, recall and F1 are calculated using a ChIP-seq gold standard for each of the individual subsets and are shown labeled as “individual”. The UpSet plot is sorted on the precision of the subsets, so that the subset with the most correct predictions is shown on the left. The same performance metrics are also calculated cumulatively (labeled as “cumulative”), starting from the subset with the highest precision (most left) and progressively adding individual subsets with the next best precision, until all regions are added together (most right). The combined subsets that result in the “max F1 iCREs” and “all iCREs” are labeled “max F1” and “all” respectively. (D) The genomic feature distribution of the “max F1 iCREs” and “all iCREs”. The total coverage in each genomic feature is given in megabases (Mb).

Figure 2

Figure 2. iCREs-based gene regulatory network (GRN) prediction and selection of drought-responsive gene sets. (A) Overview of the iCREs-based GRN inference framework, where the iCREs associated to a set of functionally-related or co-expressed genes are submitted to MINI-AC to perform motif enrichment and GRN inference. (B) On the left, there is the inverse cumulative distribution of the number of up- and down-

regulated genes (y-axis) that are common in a specific number of contrasts (x-axis). On the right, there is the activation and repression consistency score distribution of the “core” drought-responsive genes. (C) UpSet plot showing the number of common and shared drought-responsive genes per tissue, sorted by the size of the intersection set. For visualization purposes, only intersection sets with 25 or more genes are shown. (D) Clustered heatmaps showing the log-transformed fold change of gene expression in the “core” (left) and “marker tissue-specific” (right) drought-responsive gene sets.

Figure 3

Figure 3. Benchmark and analysis of iCREs-based and drought-responsive gene regulatory networks (GRNs). (A) Comparison of the performance of the “promoter-based” and “iCREs-based” GRN inference frameworks, using two different iCREs sets (maxF1 and all). The two performance metrics (F1 and area under the precision recall curve) measure how well each method predicts differentially expressed (DE) regulators given a set of tissue-specific drought-responsive genes (y-axis). (B) Clustered heatmap showing the motif enrichment rank of different DE transcription factors (TFs) (y-axis) given sets of iCREs that are associated with tissue-specific (x-axis) up-regulated genes in response to drought. The row annotations correspond to the activation and repression consistency score bins and the TF family. For visualization purposes, only TFs with an expression consistency score higher than 0.513 are shown. (C) Clustered heatmap displaying the -log10(q-value) of gene ontology (GO; x-axis) enrichment among the predicted target genes of up- and down-regulated TFs in response to drought in leaf (y-axis). The row annotations indicate the activation and repression consistency score of the TFs, and the TF family. The column annotation corresponds to the gene set (up- or down-regulated marker genes in response to drought in leaf) that yielded the indicated GO enrichment, and the number of target genes in that GO category. For visualization purposes, the enriched GO terms with 5 or fewer target genes were discarded, as well as TFs with 3 or less GO terms enriched.

Figure 4

Figure 4. Analysis of the regulatory role of transposable elements (TEs) in maize using iCREs. (A) Bar plot that shows the enrichment fold (x-axis) and q-value of TE

superfamilies and repetitive element types (y-axis) in the iCREs. (B) Clustered heatmap displaying the -log10(q-value) of gene ontology (GO; y-axis) enrichment among the genes associated with TEs within iCREs of three tandem inverted repeats (TIR) superfamilies (x-axis). The row annotations indicate the number of TE-associated genes in the corresponding GO category. The heatmap was filtered for TF-GO pairs with a q-value < 0.0008 with at least 100 TE-associated genes within that GO category. (C) Chromatin states (CS; y-axis) enrichment of different sets of genomic regions (x-axis), including all TEs and repetitive elements, iCREs, and TEs within and outside iCREs. The tables adjacent to the x-axis indicate the chromatin marks and genomic locations characteristic of each CS. (D) Box plot displaying the expression level distribution in transcripts per million (y-axis) of the genes associated with TEs within iCREs, outside iCREs, and both, binned per distance to the corresponding genes. (E) Bubble plot showing the motif enrichment rank and π-value (−log10(p-value) × enrichment fold) of different transcription factor (TF) families on the TEs within iCREs that are associated to genes annotated with “defense response to bacterium”, “regulation of flower development”, and “response to salt stress”, and that belong to the superfamilies Tc1/Mariner and PIF/Harbinger. Only the families with at least one motif with motif enrichment rank of 30 and less are shown.

Figure S1

Figure S1. Length distribution, genome coverage, and genomic feature distribution of the collected datasets with putative cis-regulatory elements (CREs), and transposable and repetitive elements. (A) The length distribution of the genomic regions in a dataset (indicated on the x-axis at the bottom), annotated with the number of regions (n) in the dataset. (B) The coverage of genomic features by the different regions of each dataset, i.e., what percentage of a given genomic feature is covered by the dataset. Datasets are annotated with their total size in megabases (Mb). Genomic features are annotated with their total size in Mb and as a percentage of the total genome size. (C) The genomic feature distribution in which the dataset regions are located, i.e., what percentage of a dataset is located in each genomic feature.

Figure S2

Figure S2. Precision-recall plot of individual CNS detection and experimental CRE-profiling methods, using a silver standard. Precision-recall plot where performance is calculated against a cross-tissue ACR dataset (A), or a UMR dataset (B).

Figure S3

Figure S3. Benchmark and comparison of the conserved non-coding sequences (CNS) detection methods. An UpSet plot showing the sizes of the full datasets on the left and the sizes of the mutually exclusive subsets on the top. Precision, recall and F1 are calculated using a ChIP-seq gold standard for each of the individual subsets and are shown labeled as “individual”. The UpSet plot is sorted on the precision of the subsets, so that the subset with the most correct predictions is shown on the left. The same performance metrics are also calculated cumulatively (labeled as “cumulative”), starting from the subset with the highest precision (most left) and progressively adding individual subsets with the next best precision, until all regions are added together (most right).

Figure S4

Figure S4. Quality control metrics of the maize drought gene expression atlas processing. (A) Box plot showing the distribution of the mapping rates of the samples per study. (B) Box plot showing the distribution of number of differentially expressed genes in the samples, per study.

Figure S5

Figure S5. Distribution of the activation and repression consistency scores of the tissue-specific “marker” drought-responsive genes.

Figure S6

Figure S6. Chromatin states (CS) enrichment per TE superfamily and repetitive elements type. Clustered heatmap showing the CS (y-axis) enrichment π-value of different TE subsets split per TE superfamily or repetitive elements type (x-axis). The

columns are annotated based on the TE superfamily/repetitive elements type, and based on the TE subset (all TEs, TEs within iCREs, and TEs outside iCREs).

Figure S7

Figure S7. Distribution of the distance to the closest gene for the indicated sets of genomic regions.

Tables

Table S1

Table S1. Performance metrics and genome coverage of individual datasets. Performance is calculated using a ChIP-seq gold standard. The adjusted p-value represents the significance of overlap between the individual dataset and the gold standard.

Table S2

Table S2. List of the maize drought studies employed in this publication, their metadata, and characteristics.

Table S3

Table S3. Gene ontology (GO) enrichment of the genes associated with different repetitive element types and transposable element (TE) superfamilies that are found within iCREs.

Table S4

Table S4. Number of, median value of the transcripts per million, and p-value of the genes associated with TEs within, outside, and within and outside iCREs. The computations were binned per distance to the closest gene.

References

Batista, Rita A, Jordi Moreno-Romero, Yichun Qiu, Joram van Boven, Juan Santos-González, Duarte D Figueiredo, and Claudia Köhler. 2019. “The MADS-Box Transcription Factor PHERES1 Controls Imprinting in the Endosperm by Binding to Domesticated Transposons.” Edited by Daniel Zilberman and Christian S

Hardtke. *eLife* 8 (December):e50541. https://doi.org/10.7554/eLife.50541.

Benjamini, Yoav, and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1): 289–300.

Berrío, Rubén Tenorio, Kevin Verstaen, Niels Vandamme, Julie Pevernagie, Ignacio Achon, Julie van Duyse, Gert van Isterdael, et al. 2022. “Single-Cell Transcriptomics Sheds Light on the Identity and Metabolism of Developing Leaf Cells.” *Plant Physiology* 188 (2): 898–918. https://doi.org/10.1093/plphys/kiab489.

Boyle, Alan P., Sean Davis, Hennady P. Shulha, Paul Meltzer, Elliott H. Margulies, Zhiping Weng, Terrence S. Furey, and Gregory E. Crawford. 2008. “High-Resolution Mapping and Characterization of Open Chromatin across the Genome.” *Cell* 132 (2): 311–22. https://doi.org/10.1016/j.cell.2007.12.014.

Bravo González-Blas, Carmen, Seppe De Winter, Gert Hulselmans, Nikolai Hecker, Irina Matetovici, Valerie Christiaens, Suresh Poovathingal, Jasper Wouters, Sara Aibar, and Stein Aerts. 2023. “SCENIC+: Single-Cell Multiomic Inference of Enhancers and Gene Regulatory Networks.” *Nature Methods*, July, 1–13. https://doi.org/10.1038/s41592-023-01938-4.

Breuer, Christian, Kengo Morohashi, Ayako Kawamura, Naoki Takahashi, Takashi Ishida, Masaaki Umeda, Erich Grotewold, and Keiko Sugimoto. 2012. “Transcriptional Repression of the APC/C Activator CCS52A1 Promotes Active Termination of Cell Growth.” *The EMBO Journal* 31 (24): 4488–4501. https://doi.org/10.1038/emboj.2012.294.

Britten, R. J., and E. H. Davidson. 1971. “Repetitive and Non-Repetitive DNA Sequences and a Speculation on the Origins of Evolutionary Novelty.” *The Quarterly Review of Biology* 46 (2): 111–38. https://doi.org/10.1086/406830.

Brooks, Matthew D, Che-Lun Juang, Manpreet Singh Katari, José M Alvarez, Angelo Pasquino, Hung-Jui Shih, Ji Huang, Carly Shanks, Jacopo Cirrone, and Gloria M Coruzzi. 2020. “ConnecTF: A Platform to Integrate Transcription Factor-Gene Interactions and Validate Regulatory Networks.” *Plant Physiology*. https://doi.org/10.1093/plphys/kiaa012/5988590.

Buenrostro, Jason D., Beijing Wu, Howard Y. Chang, and William J. Greenleaf. 2015. “ATAC-Seq: A Method for Assaying Chromatin Accessibility Genome-Wide.” *Current Protocols in Molecular Biology* 109 (January):21.29.1-21.29.9. https://doi.org/10.1002/0471142727.mb2129s109.

Burgess, Steven J., Ivan Reyna-Llorens, Sean R. Stevenson, Pallavi Singh, Katja Jaeger, and Julian M. Hibberd. 2019. “Genome-Wide Transcription Factor Binding in Leaves from C3 and C4 Grasses.” *The Plant Cell* 31 (10): 2297–2314. https://doi.org/10.1105/TPC.19.00078.

Caro, Elena, Bénédicte Desvoyes, and Crisanto Gutierrez. 2012. “GTL1 Keeps Cell Growth and Nuclear Ploidy under Control.” *The EMBO Journal* 31 (24): 4483–85. https://doi.org/10.1038/emboj.2012.311.

Chen, Dijun, Wenhao Yan, Liang Yu Fu, and Kerstin Kaufmann. 2018. “Architecture of Gene Regulatory Networks Controlling Flower Development in Arabidopsis Thaliana.” *Nature Communications* 9 (1): 1–13. https://doi.org/10.1038/s41467-018-06772-3.

Chen, Xujun, Cheng Li, Han Wang, and Zejian Guo. 2019. “WRKY Transcription Factors: Evolution, Binding, and Action.” *Phytopathology Research* 1 (1): 13. https://doi.org/10.1186/s42483-019-0022-x.

Clark, Natalie M., Trevor M. Nolan, Ping Wang, Gaoyuan Song, Christian Montes,

Conner T. Valentine, Hongqing Guo, Rosangela Sozzani, Yanhai Yin, and Justin W. Walley. 2021. “Integrated Omics Networks Reveal the Temporal Signaling Events of Brassinosteroid Response in Arabidopsis.” *Nature Communications* 12 (1): 5858. https://doi.org/10.1038/s41467-021-26165-3.

Cooper, Gregory M., and Christopher D. Brown. 2008. “Qualifying the Relationship between Sequence Conservation and Molecular Function.” *Genome Research* 18 (2): 201–5. https://doi.org/10.1101/gr.7205808.

Crisp, Peter A., Diep R. Ganguly, Aaron B. Smith, Kevin D. Murray, Gonzalo M. Estavillo, Iain Searle, Ethan Ford, et al. 2017. “Rapid Recovery Gene Downregulation during Excess-Light Stress and Recovery in Arabidopsis.” *The Plant Cell* 29 (8): 1836–63. https://doi.org/10.1105/tpc.16.00828.

Crisp, Peter A., Alexandre P. Marand, Jaclyn M. Noshay, Peng Zhou, Zefu Lu, Robert J. Schmitz, and Nathan M. Springer. 2020. “Stable Unmethylated DNA Demarcates Expressed Genes and Their Cis-Regulatory Space in Plant Genomes.” *Proceedings of the National Academy of Sciences of the United States of America* 117 (38): 23991–0. https://doi.org/10.1073/pnas.2010250117.

Davydov, Eugene V., David L. Goode, Marina Sirota, Gregory M. Cooper, Arend Sidow, and Serafim Batzoglou. 2010. “Identifying a High Fraction of the Human Genome to Be under Selective Constraint Using GERP++.” *PLoS Computational Biology* 6 (12): e1001025. https://doi.org/10.1371/journal.pcbi.1001025.

De Clercq, Inge, Jan Van de Velde, Xiaopeng Luo, Li Liu, Veronique Storme, Michiel Van Bel, Robin Pottie, Dries Vaneechoutte, Frank Van Breusegem, and Klaas Vandepoele. 2021. “Integrative Inference of Transcriptional Networks in Arabidopsis Yields Novel ROS Signalling Regulators.” *Nature Plants 2021 7:4* 7 (4): 500–513. https://doi.org/10.1038/s41477-021-00894-1.

De Witte, Dieter, Jan Van De Velde, Dries Decap, Michiel Van Bel, Pieter Audenaert, Piet Demeester, Bart Dhoedt, Klaas Vandepoele, and Jan Fostier. 2015. “BLSSpeller: Exhaustive Comparative Discovery of Conserved Cis-Regulatory Elements.” *Bioinformatics* 31 (23): 3758–66. https://doi.org/10.1093/bioinformatics/btv466.

Deneweth, Jan, Yves Van de Peer, and Vanessa Vermeirssen. 2022. “Nearby Transposable Elements Impact Plant Stress Gene Regulatory Networks: A Meta-Analysis in A. Thaliana and S. Lycopersicum.” *BMC Genomics* 23 (1): 18. https://doi.org/10.1186/s12864-021-08215-8.

Depuydt, Thomas, Bert De Rybel, and Klaas Vandepoele. 2023. “Charting Plant Gene Functions in the Multi-Omics and Single-Cell Era.” *Trends in Plant Science* 28 (3): 283–96. https://doi.org/10.1016/j.tplants.2022.09.008.

Ding, Ke, Shanwen Sun, Yang Luo, Chaoyue Long, Jingwen Zhai, Yixiao Zhai, and Guohua Wang. 2022. “PlantCADB: A Comprehensive Plant Chromatin Accessibility Database.” *Genomics, Proteomics & Bioinformatics*, October. https://doi.org/10.1016/j.gpb.2022.10.005.

Ding, Pingtao, Toshiyuki Sakai, Ram Krishna Shrestha, Nicolas Manosalva Perez, Wenbin Guo, Bruno Pok Man Ngou, Shengbo He, et al. 2021. “Chromatin Accessibility Landscapes Activated by Cell-Surface and Intracellular Immune Receptors.” Edited by John Lunn. *Journal of Experimental Botany* 72 (22): 7927–41. https://doi.org/10.1093/jxb/erab373.

Dong, Pengfei, Xiaoyu Tu, Po-Yu Chu, Peitao Lü, Ning Zhu, Donald Grierson, Baijuan Du, Pinghua Li, and Silin Zhong. 2017. “3D Chromatin Architecture of Large Plant Genomes Determined by Local A/B Compartments.” *Molecular Plant* 10 (12): 1497–1509. https://doi.org/10.1016/j.molp.2017.11.005.

Dubois, Marieke, and Dirk Inzé. 2020. “Plant Growth under Suboptimal Water Conditions: Early Responses and Methods to Study Them.” *Journal of Experimental Botany* 71 (5): 1706–22. https://doi.org/10.1093/jxb/eraa037.

Eichten, Steven R., and Nathan M. Springer. 2015. “Minimal Evidence for Consistent Changes in Maize DNA Methylation Patterns Following Environmental Stress.” *Frontiers in Plant Science* 6. https://www.frontiersin.org/articles/10.3389/fpls.2015.00308.

Eichten, Steven R., Matthew W. Vaughn, Peter J. Hermanson, and Nathan M. Springer. 2013. “Variation in DNA Methylation Patterns Is More Common among Maize Inbreds than among Tissues.” *The Plant Genome* 6 (2): plantgenome2012.06.0009. https://doi.org/10.3835/plantgenome2012.06.0009.

Eli, Rodgers Melnick, L. Vera Daniel, W. Bass Hank, and S. Buckler Edward. 2016. “Open Chromatin Reveals the Functional Maize Genome.” *Proceedings of the National Academy of Sciences of the United States of America* 113 (22): 3177–84. https://doi.org/10.1073/pnas.1525244113.

Emms, David M., and Steven Kelly. 2019. “OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics.” *Genome Biology* 20 (1): 238. https://doi.org/10.1186/s13059-019-1832-y.

Ewels, Philip A., Alexander Peltzer, Sven Fillinger, Harshil Patel, Johannes Alneberg, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso, and Sven Nahnsen. 2020. “The Nf-Core Framework for Community-Curated Bioinformatics Pipelines.” *Nature Biotechnology* 38 (3): 276–78. https://doi.org/10.1038/s41587-020-0439-x.

Ferrari, Camilla, Nicolás Manosalva Pérez, and Klaas Vandepoele. 2022. “MINI-EX: Integrative Inference of Single-Cell Gene Regulatory Networks in Plants.” *Molecular Plant* 15 (11): 1807–24. https://doi.org/10.1016/j.molp.2022.10.016.

Feschotte, Cédric. 2008. “Transposable Elements and the Evolution of Regulatory Networks.” *Nature Reviews Genetics* 9 (5): 397–405. https://doi.org/10.1038/nrg2337.

Fleck, Jonas Simon, Sophie Martina Johanna Jansen, Damian Wollny, Fides Zenk, Makiko Seimiya, Akanksha Jain, Ryoko Okamoto, et al. 2023. “Inferring and Perturbing Cell Fate Regulomes in Human Brain Organoids.” *Nature* 621 (7978): 365–72. https://doi.org/10.1038/s41586-022-05279-8.

Fujimori, Toru, Takafumi Yamashino, Takahiko Kato, and Takeshi Mizuno. 2004. “Circadian-Controlled Basic/Helix-Loop-Helix Factor, PIL6, Implicated in Light-Signal Transduction in Arabidopsis Thaliana.” *Plant and Cell Physiology* 45 (8): 1078–86. https://doi.org/10.1093/pcp/pch124.

Fullwood, Melissa J., Mei Hui Liu, You Fu Pan, Jun Liu, Han Xu, Yusoff Bin Mohamed, Yuriy L. Orlov, et al. 2009. “An Oestrogen-Receptor-α-Bound Human Chromatin Interactome.” *Nature* 462 (7269): 58–64. https://doi.org/10.1038/nature08497.

Ganguly, Diep R., Peter A. Crisp, Steven R. Eichten, and Barry J. Pogson. 2017. “The Arabidopsis DNA Methylome Is Stable under Transgenerational Drought Stress.” *Plant Physiology* 175 (4): 1893–1912. https://doi.org/10.1104/pp.17.00744.

Gaudinier, Allison, Joel Rodriguez-Medina, Lifang Zhang, Andrew Olson, Christophe Liseron-Monfils, Anne Maarit Bågman, Jessica Foret, et al. 2018. “Transcriptional Regulation of Nitrogen-Associated Metabolism and Growth.” *Nature* 563 (7730): 259–64. https://doi.org/10.1038/s41586-018-0656-3.

Grau, Joaquín, and José M. Franco-Zorrilla. 2022. “TDTHub, a Web Server Tool for the Analysis of Transcription Factor Binding Sites in Plants.” *The Plant Journal* 111 (4): 1203–15. https://doi.org/10.1111/tpj.15873.

Gulzar, Faiza, Jingye Fu, Chenying Zhu, Jie Yan, Xinglin Li, Tehseen Ahmad Meraj, Qinqin Shen, Beenish Hassan, and Qiang Wang. 2021. “Maize WRKY Transcription Factor ZmWRKY79 Positively Regulates Drought Tolerance through Elevating ABA Biosynthesis.” *International Journal of Molecular Sciences* 22 (18): 10080. https://doi.org/10.3390/ijms221810080.

Haudry, Annabelle, Adrian E Platts, Emilio Vello, Douglas R Hoen, Mickael Leclercq, Robert J Williamson, Ewa Forczek, et al. 2013. “An Atlas of over 90,000 Conserved Noncoding Sequences Provides Insight into Crucifer Regulatory Regions.” *Nature Genetics* 45 (8): 891–98. https://doi.org/10.1038/ng.2684.

Hinrichs, A. S., D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, et al. 2006. “The UCSC Genome Browser Database: Update 2006.” *Nucleic Acids Research* 34 (suppl_1): D590–98. https://doi.org/10.1093/nar/gkj144.

Hirsch, Cory D., and Nathan M. Springer. 2017. “Transposable Element Influences on Gene Expression in Plants.” *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, Plant Gene Regulatory Mechanisms and Networks, 1860 (1): 157–65. https://doi.org/10.1016/j.bbagrm.2016.05.010.

Jiang, Junyao, Pin Lyu, Jinlian Li, Sunan Huang, Jiawang Tao, Seth Blackshaw, Jiang Qian, and Jie Wang. 2022. “IReNA: Integrated Regulatory Network Analysis of Single-Cell Transcriptomes and Chromatin Accessibility Profiles.” *iScience* 25 (11). https://doi.org/10.1016/j.isci.2022.105359.

Jiao, Yinping, Paul Peluso, Jinghua Shi, Tiffany Liang, Michelle C. Stitzer, Bo Wang, Michael S. Campbell, et al. 2017. “Improved Maize Reference Genome with Single-Molecule Technologies.” *Nature* 546 (7659): 524–27. https://doi.org/10.1038/nature22971.

Kajala, Kaisa, Mona Gouran, Lidor Shaar-Moshe, G. Alex Mason, Joel Rodriguez-Medina, Dorota Kawa, Germain Pauluzzi, et al. 2021. “Innovation, Conservation, and Repurposing of Gene Function in Root Cell Type Development.” *Cell* 184 (12): 3333-3348.e19. https://doi.org/10.1016/j.cell.2021.04.024.

Kamimoto, Kenji, Blerta Stringa, Christy M. Hoffmann, Kunal Jindal, Lilianna Solnica-Krezel, and Samantha A. Morris. 2023. “Dissecting Cell Identity via Network Inference and in Silico Gene Perturbation.” *Nature* 614 (7949): 742–51. https://doi.org/10.1038/s41586-022-05688-9.

Kawakatsu, Taiji, Tim Stuart, Manuel Valdes, Natalie Breakfield, Robert J. Schmitz, Joseph R. Nery, Mark A. Urich, et al. 2016. “Unique Cell-Type-Specific Patterns of DNA Methylation in the Root Meristem.” *Nature Plants* 2 (5). https://doi.org/10.1038/NPLANTS.2016.58.

Kent, W. James, Robert Baertsch, Angie Hinrichs, Webb Miller, and David Haussler. 2003. “Evolution’s Cauldron: Duplication, Deletion, and Rearrangement in the Mouse and Human Genomes.” *Proceedings of the National Academy of Sciences* 100 (20): 11484–89. https://doi.org/10.1073/pnas.1932072100.

Kimura, M. 1991. “Recent Development of the Neutral Theory Viewed from the Wrightian Tradition of Theoretical Population Genetics.” *Proceedings of the National Academy of Sciences* 88 (14): 5969–73. https://doi.org/10.1073/pnas.88.14.5969.

Kulkarni, Shubhada R., Dries Vaneechoutte, Jan Van de Velde, and Klaas Vandepoele. 2018. “TF2Network: Predicting Transcription Factor Regulators and Gene Regulatory Networks in Arabidopsis Using Publicly Available Binding Site Information.” *Nucleic Acids Research* 46 (6): e31. https://doi.org/10.1093/nar/gkx1279.

Kunarso, Galih, Na-Yu Chia, Justin Jeyakani, Catalina Hwang, Xinyi Lu, Yun-Shen

Chan, Huck-Hui Ng, and Guillaume Bourque. 2010. “Transposable Elements Have Rewired the Core Regulatory Network of Human Embryonic Stem Cells.” *Nature Genetics* 42 (7): 631–34. https://doi.org/10.1038/ng.600.

Lai, Xuelei, Romain Blanc-Mathieu, Loïc GrandVuillemin, Ying Huang, Arnaud Stigliani, Jérémy Lucas, Emmanuel Thévenon, et al. 2021. “The LEAFY Floral Regulator Displays Pioneer Transcription Factor Properties.” *Molecular Plant* 14 (5): 829–37. https://doi.org/10.1016/j.molp.2021.03.004.

Le, Tuan-Ngoc, Ulrike Schumann, Neil A. Smith, Sameer Tiwari, Phil Chi Khang Au, Qian-Hao Zhu, Jennifer M. Taylor, et al. 2014. “DNA Demethylases Target Promoter Transposable Elements to Positively Regulate Stress Responsive Genes in Arabidopsis.” *Genome Biology* 15 (9): 458. https://doi.org/10.1186/s13059-014-0458-3.

Li, En, Han Liu, Liangliang Huang, Xiangbo Zhang, Xiaomei Dong, Weibin Song, Haiming Zhao, and Jinsheng Lai. 2019. “Long-Range Interactions between Proximal and Distal Regulatory Regions in Maize.” *Nature Communications* 10 (1): 2633. https://doi.org/10.1038/s41467-019-10603-4.

Li, Xing, Yong Wang, Chengcheng Cai, Jialei Ji, Fengqing Han, Lei Zhang, Shumin Chen, et al. 2024. “Large-Scale Gene Expression Alterations Introduced by Structural Variation Drive Morphotype Diversification in Brassica Oleracea.” *Nature Genetics*, February, 1–13. https://doi.org/10.1038/s41588-024-01655-4.

Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. “Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome.” *Science* 326 (5950): 289–93. https://doi.org/10.1126/science.1181369.

Liu, Yue, Tian Tian, Kang Zhang, Qi You, Hengyu Yan, Nannan Zhao, Xin Yi, Wenying Xu, and Zhen Su. 2018. “PCSD: A Plant Chromatin State Database.” *Nucleic Acids Research* 46 (D1): D1157–67. https://doi.org/10.1093/nar/gkx919.

Lu, Dandan, Ting Wang, Staffan Persson, Bernd Mueller-Roeber, and Jos H. M. Schippers. 2014. “Transcriptional Control of ROS Homeostasis by KUODA1 Regulates Cell Expansion during Leaf Development.” *Nature Communications* 5 (1): 3767. https://doi.org/10.1038/ncomms4767.

Lu, Zefu, Brigitte T. Hofmeister, Christopher Vollmers, Rebecca M. DuBois, and Robert J. Schmitz. 2017. “Combining ATAC-Seq with Nuclei Sorting for Discovery of Cis-Regulatory Regions in Plant Genomes.” *Nucleic Acids Research* 45 (6): e41. https://doi.org/10.1093/nar/gkw1179.

Lu, Zefu, Alexandre P. Marand, William A. Ricci, Christina L. Ethridge, Xiaoyu Zhang, and Robert J. Schmitz. 2019. “The Prevalence, Evolution and Chromatin Signatures of Plant Regulatory Elements.” *Nature Plants* 5 (12): 1250–59. https://doi.org/10.1038/s41477-019-0548-z.

Lynch, Vincent J., Robert D. Leclerc, Gemma May, and Günter P. Wagner. 2011. “Transposon-Mediated Rewiring of Gene Regulatory Networks Contributed to the Evolution of Pregnancy in Mammals.” *Nature Genetics* 43 (11): 1154–59. https://doi.org/10.1038/ng.917.

Maher, Kelsey A., Marko Bajic, Kaisa Kajala, Mauricio Reynoso, Germain Pauluzzi, Donnelly A. West, Kristina Zumstein, et al. 2018. “Profiling of Accessible Chromatin Regions across Multiple Plant Species and Cell Types Reveals Common Gene Regulatory Principles and New Control Modules.” *Plant Cell* 30 (1): 15–36. https://doi.org/10.1105/tpc.17.00581.

Makarevitch, Irina, Amanda J. Waters, Patrick T. West, Michelle Stitzer, Candice N. Hirsch, Jeffrey Ross-Ibarra, and Nathan M. Springer. 2015. “Transposable

Elements Contribute to Activation of Maize Genes in Response to Abiotic Stress." *PLOS Genetics* 11 (1): e1004915. https://doi.org/10.1371/journal.pgen.1004915.

Manosalva Pérez, Nicolás, Camilla Ferrari, Julia Engelhorn, Thomas Depuydt, Hilde Nelissen, Thomas Hartwig, and Klaas Vandepoele. 2024. "MINI-AC: Inference of Plant Gene Regulatory Networks Using Bulk or Single-Cell Accessible Chromatin Profiles." *The Plant Journal* 117 (1): 280–301. https://doi.org/10.1111/tpj.16483.

Mapleson, Daniel, Gonzalo Garcia Accinelli, George Kettleborough, Jonathan Wright, and Bernardo J Clavijo. 2017. "KAT: A K-Mer Analysis Toolkit to Quality Control NGS Datasets and Genome Assemblies." *Bioinformatics* 33 (4): 574–76. https://doi.org/10.1093/bioinformatics/btw663.

Marand, Alexandre P, Zongliang Chen, Andrea Gallavotti, and Robert J Schmitz. 2021. "A Cis-Regulatory Atlas in Maize at Single-Cell Resolution." *Cell* 184 (11): 3041-3055.e21. https://doi.org/10.1016/j.cell.2021.04.014.

Marand, Alexandre P., Andrea L. Eveland, Kerstin Kaufmann, and Nathan M. Springer. 2023. "Cis-Regulatory Elements in Plant Development, Adaptation, and Evolution." *Annual Review of Plant Biology* 74 (1): 111–37. https://doi.org/10.1146/annurev-arplant-070122-030236.

Martin, Fergal J, M Ridwan Amode, Alisha Aneja, Olanrewaju Austine-Orimoloye, Andrey G Azov, If Barnes, Arne Becker, et al. 2023. "Ensembl 2023." *Nucleic Acids Research* 51 (D1): D933–41. https://doi.org/10.1093/nar/gkac958.

Noll, Markus. 1974. "Subunit Structure of Chromatin." *Nature* 251 (5472): 249–51. https://doi.org/10.1038/251249a0.

Noshay, Jaclyn M, Alexandre P Marand, Sarah N Anderson, Peng Zhou, Maria Katherine Mejia Guerra, Zefu Lu, Christine H O'Connor, et al. 2021. "Assessing the Regulatory Potential of Transposable Elements Using Chromatin Accessibility Profiles of Maize Transposons." Edited by K Bomblies. *Genetics* 217 (1): 1–13. https://doi.org/10.1093/genetics/iyaa003.

Oka, Rurika, Johan Zicola, Blaise Weber, Sarah N. Anderson, Charlie Hodgman, Jonathan I. Gent, Jan Jaap Wesselink, et al. 2017. "Genome-Wide Mapping of Transcriptional Enhancer Candidates Using DNA and Chromatin Features in Maize." *Genome Biology* 18 (1): 1–24. https://doi.org/10.1186/s13059-017-1273-4.

Oudet, P., M. Gross-Bellard, and P. Chambon. 1975. "Electron Microscopic and Biochemical Evidence That Chromatin Structure Is a Repeating Unit." *Cell* 4 (4): 281–300. https://doi.org/10.1016/0092-8674(75)90149-X.

Patro, Rob, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. 2017. "Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression." *Nature Methods* 14 (4): 417–19. https://doi.org/10.1038/nmeth.4197.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12 (85): 2825–30.

Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26 (6): 841–42. https://doi.org/10.1093/bioinformatics/btq033.

Raines, Tracy, Carly Shanks, Chia-Yi Cheng, Duncan McPherson, Cristiana T. Argueso, Hyo J. Kim, José M. Franco-Zorrilla, et al. 2016. "The Cytokinin Response Factors Modulate Root and Shoot Growth and Promote Leaf Senescence in Arabidopsis." *The Plant Journal* 85 (1): 134–47. https://doi.org/10.1111/tpj.13097.

Rebollo, Rita, Mark T. Romanish, and Dixie L. Mager. 2012. "Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes."

Annual Review of Genetics 46 (1): 21–42. https://doi.org/10.1146/annurev-genet-110711-155621.

Reynoso, Mauricio A., Kaisa Kajala, Marko Bajic, Donnelly A. West, Germain Pauluzzi, Andrew I. Yao, Kathryn Hatch, et al. 2019. "Evolutionary Flexibility in Flooding Response Circuitry in Angiosperms." *Science* 365 (6459): 1291–95. https://doi.org/10.1126/science.aax8862.

Ricci, William A., Zefu Lu, Lexiang Ji, Alexandre P. Marand, Christina L. Ethridge, Nathalie G. Murphy, Jaclyn M. Noshay, et al. 2019. "Widespread Long-Range Cis-Regulatory Elements in the Maize Genome." *Nature Plants* 5 (12): 1237–49. https://doi.org/10.1038/s41477-019-0547-0.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2010. "edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics* 26 (1): 139–40. https://doi.org/10.1093/bioinformatics/btp616.

Roquis, David, Marta Robertson, Liang Yu, Michael Thieme, Magdalena Julkowska, and Etienne Bucher. 2021. "Genomic Impact of Stress-Induced Transposable Element Mobility in Arabidopsis." *Nucleic Acids Research* 49 (18): 10431–47. https://doi.org/10.1093/nar/gkab828.

Savadel, Savannah D., Thomas Hartwig, Zachary M. Turpin, Daniel L. Vera, Pei-Yau Lung, Xin Sui, Max Blank, et al. 2021. "The Native Cistrome and Sequence Motif Families of the Maize Ear." Edited by Nathan M. Springer. *PLOS Genetics* 17 (8): e1009689. https://doi.org/10.1371/journal.pgen.1009689.

Schmitz, Robert J., Erich Grotewold, and Maike Stam. 2022. "Cis-Regulatory Sequences in Plants: Their Importance, Discovery, and Future Challenges." *The Plant Cell* 34 (2): 718–41. https://doi.org/10.1093/PLCELL/KOAB281.

Siepel, Adam, Gill Bejerano, Jakob S. Pedersen, Angie S. Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, et al. 2005. "Evolutionarily Conserved Elements in Vertebrate, Insect, Worm, and Yeast Genomes." *Genome Research* 15 (8): 1034–50. https://doi.org/10.1101/gr.3715005.

Song, Baoxing, Edward S. Buckler, Hai Wang, Yaoyao Wu, Evan Rees, Elizabeth A. Kellogg, Daniel J. Gates, et al. 2021. "Conserved Noncoding Sequences Provide Insights into Regulatory Sequence and Loss of Gene Expression in Maize." *Genome Research* 31 (7): 1245–57. https://doi.org/10.1101/gr.266528.120.

Soppe, Wim J. J., Zuzana Jasencakova, Andreas Houben, Tetsuji Kakutani, Armin Meister, Michael S. Huang, Steven E. Jacobsen, Ingo Schubert, and Paul F. Fransz. 2002. "DNA Methylation Controls Histone H3 Lysine 9 Methylation and Heterochromatin Assembly in Arabidopsis." *The EMBO Journal* 21 (23): 6549–59. https://doi.org/10.1093/emboj/cdf657.

Strader, Lucia, Dolf Weijers, and Doris Wagner. 2022. "Plant Transcription Factors — Being in the Right Place with the Right Company." *Current Opinion in Plant Biology* 65 (February):102136. https://doi.org/10.1016/j.pbi.2021.102136.

Sullivan, Alessandra M., Andrej A. Arsovski, Janne Lempe, Kerry L. Bubb, Matthew T. Weirauch, Peter J. Sabo, Richard Sandstrom, et al. 2014. "Mapping and Dynamics of Regulatory DNA and Transcription Factor Networks in A. Thaliana." *Cell Reports* 8 (6): 2015–30. https://doi.org/10.1016/j.celrep.2014.08.019.

Sun, Yonghao, Liang Dong, Ying Zhang, Da Lin, Weize Xu, Changxiong Ke, Linqian Han, et al. 2020. "3D Genome Architecture Coordinates Trans and Cis Regulation of Differentially Expressed Ear and Tassel Genes in Maize." *Genome Biology* 21 (1): 143. https://doi.org/10.1186/s13059-020-02063-7.

Tian, Feng, De Chang Yang, Yu Qi Meng, Jinpu Jin, and Ge Gao. 2020. "PlantRegMap:

Charting Functional Regulatory Maps in Plants." *Nucleic Acids Research* 48 (D1): D1104–13. https://doi.org/10.1093/nar/gkz1020.

Tu, Xiaoyu, María Katherine Mejía-Guerra, Jose A Valdes Franco, David Tzeng, Po Yu Chu, Wei Shen, Yingying Wei, et al. 2020. "Reconstructing the Maize Leaf Regulatory Network Using ChIP-Seq Data of 104 Transcription Factors." *Nature Communications* 11 (1). https://doi.org/10.1038/s41467-020-18832-8.

Van De Velde, Jan, Ken S. Heyndrickx, and Klaas Vandepoele. 2014. "Inference of Transcriptional Networks in Arabidopsis through Conserved Noncoding Sequence Analysis." *Plant Cell* 26 (7): 2729–45. https://doi.org/10.1105/tpc.114.127001.

Van Bel, Michiel, Francesca Silvestri, Eric M Weitz, Lukasz Kreft, Alexander Botzki, Frederik Coppens, and Klaas Vandepoele. 2022. "PLAZA 5.0: Extending the Scope and Power of Comparative and Functional Genomics in Plants." *Nucleic Acids Research* 50 (D1): D1468–74. https://doi.org/10.1093/nar/gkab1024.

Vaneechoutte, Dries, and Klaas Vandepoele. 2019. "Curse: Building Expression Atlases and Co-Expression Networks from Public RNA-Seq Data." *Bioinformatics* 35 (16): 2880–81. https://doi.org/10.1093/bioinformatics/bty1052.

Wang, Chang-Tao, Jing-Na Ru, Yong-Wei Liu, Jun-Feng Yang, Meng Li, Zhao-Shi Xu, and Jin-Dong Fu. 2018. "The Maize WRKY Transcription Factor ZmWRKY40 Confers Drought Resistance in Transgenic Arabidopsis." *International Journal of Molecular Sciences* 19 (9): 2580. https://doi.org/10.3390/ijms19092580.

Wang, Ting, Jue Zeng, Craig B. Lowe, Robert G. Sellers, Sofie R. Salama, Min Yang, Shawn M. Burgess, Rainer K. Brachmann, and David Haussler. 2007. "Species-Specific Endogenous Retroviruses Shape the Transcriptional Network of the Human Tumor Suppressor Protein P53." *Proceedings of the National Academy of Sciences* 104 (47): 18613–18. https://doi.org/10.1073/pnas.0703637104.

Wang, Xin, Lu Bai, Gene O. Bryant, and Mark Ptashne. 2011. "Nucleosomes and the Accessibility Problem." *Trends in Genetics* 27 (12): 487–92. https://doi.org/10.1016/j.tig.2011.09.001.

Weintraub, Harold, and Mark Groudine. 1976. "Chromosomal Subunits in Active Genes Have an Altered Conformation." *Science* 193 (4256): 848–56. https://doi.org/10.1126/science.948749.

Wicker, Thomas, François Sabot, Aurélie Hua-Van, Jeffrey L. Bennetzen, Pierre Capy, Boulos Chalhoub, Andrew Flavell, et al. 2007. "A Unified Classification System for Eukaryotic Transposable Elements." *Nature Reviews Genetics* 8 (12): 973–82. https://doi.org/10.1038/nrg2165.

Wilkins, Olivia, Christoph Hafemeister, Anne Plessis, Meisha-Marika Holloway-Phillips, Gina M. Pham, Adrienne B. Nicotra, Glenn B. Gregorio, et al. 2016. "EGRINs (Environmental Gene Regulatory Influence Networks) in Rice That Function in the Response to Water Deficit, High Temperature, and Agricultural Environments." *The Plant Cell* 28 (10): 2365–84. https://doi.org/10.1105/tpc.16.00158.

Woodhouse, Margaret R., Ethalinda K. Cannon, John L. Portwood, Lisa C. Harper, Jack M. Gardiner, Mary L. Schaeffer, and Carson M. Andorf. 2021. "A Pan-Genomic Approach to Genome Databases Using Maize as a Model System." *BMC Plant Biology* 21 (1): 385. https://doi.org/10.1186/s12870-021-03173-5.

Wu, Yaoyao, Lynn Johnson, Baoxing Song, Cinta Romay, Michelle Stitzer, Adam Siepel, Edward Buckler, and Armin Scheben. 2022. "A Multiple Alignment Workflow Shows the Effect of Repeat Masking and Parameter Tuning on Alignment in Plants." *The Plant Genome* 15 (2): e20204. https://doi.org/10.1002/tpg2.20204.

Yoo, Chan Yul, Heather E. Pence, Jing Bo Jin, Kenji Miura, Michael J. Gosney, Paul M. Hasegawa, and Michael V. Mickelbart. 2010. “The Arabidopsis GTL1 Transcription Factor Regulates Water Use Efficiency and Drought Tolerance by Modulating Stomatal Density via Transrepression of SDD1.” *The Plant Cell* 22 (12): 4128–41. https://doi.org/10.1105/tpc.110.078691.

Yuan, Guo-Cheng, Yuen-Jong Liu, Michael F. Dion, Michael D. Slack, Lani F. Wu, Steven J. Altschuler, and Oliver J. Rando. 2005. “Genome-Scale Identification of Nucleosome Positions in S. Cerevisiae.” *Science* 309 (5734): 626–30. https://doi.org/10.1126/science.1112178.

Zander, Mark, Mathew G. Lewsey, Natalie M. Clark, Lingling Yin, Anna Bartlett, J. Paola Saldierna Guzmán, Elizabeth Hann, et al. 2020. “Integrated Multi-Omics Framework of the Plant Response to Jasmonic Acid.” *Nature Plants* 6 (3): 290–302. https://doi.org/10.1038/s41477-020-0605-7.

Zhao, Hainan, Wenli Zhang, Lifen Chen, Lei Wang, Alexandre P. Marand, Yufeng Wu, and Jiming Jiang. 2018. “Proliferation of Regulatory DNA Elements Derived from Transposable Elements in the Maize Genome.” *Plant Physiology* 176 (4): 2789–2803. https://doi.org/10.1104/pp.17.01467.

Zhao, Lili, Jingwei Yan, Yang Xiang, Yue Sun, and Aying Zhang. 2021. “ZmWRKY104 Transcription Factor Phosphorylated by ZmMPK6 Functioning in ABA-Induced Antioxidant Defense and Enhance Drought Tolerance in Maize.” *Biology* 10 (9): 893. https://doi.org/10.3390/biology10090893.

Zhou, Peng, Zhi Li, Erika Magnusson, Fabio Gomez Cano, Peter A. Crisp, Jaclyn M. Noshay, Erich Grotewold, Candice N. Hirsch, Steven P. Briggs, and Nathan M. Springer. 2020. “Meta Gene Regulatory Networks in Maize Highlight Functionally Relevant Regulatory Interactions.” *Plant Cell* 32 (5): 1377–96. https://doi.org/10.1105/tpc.20.00080.

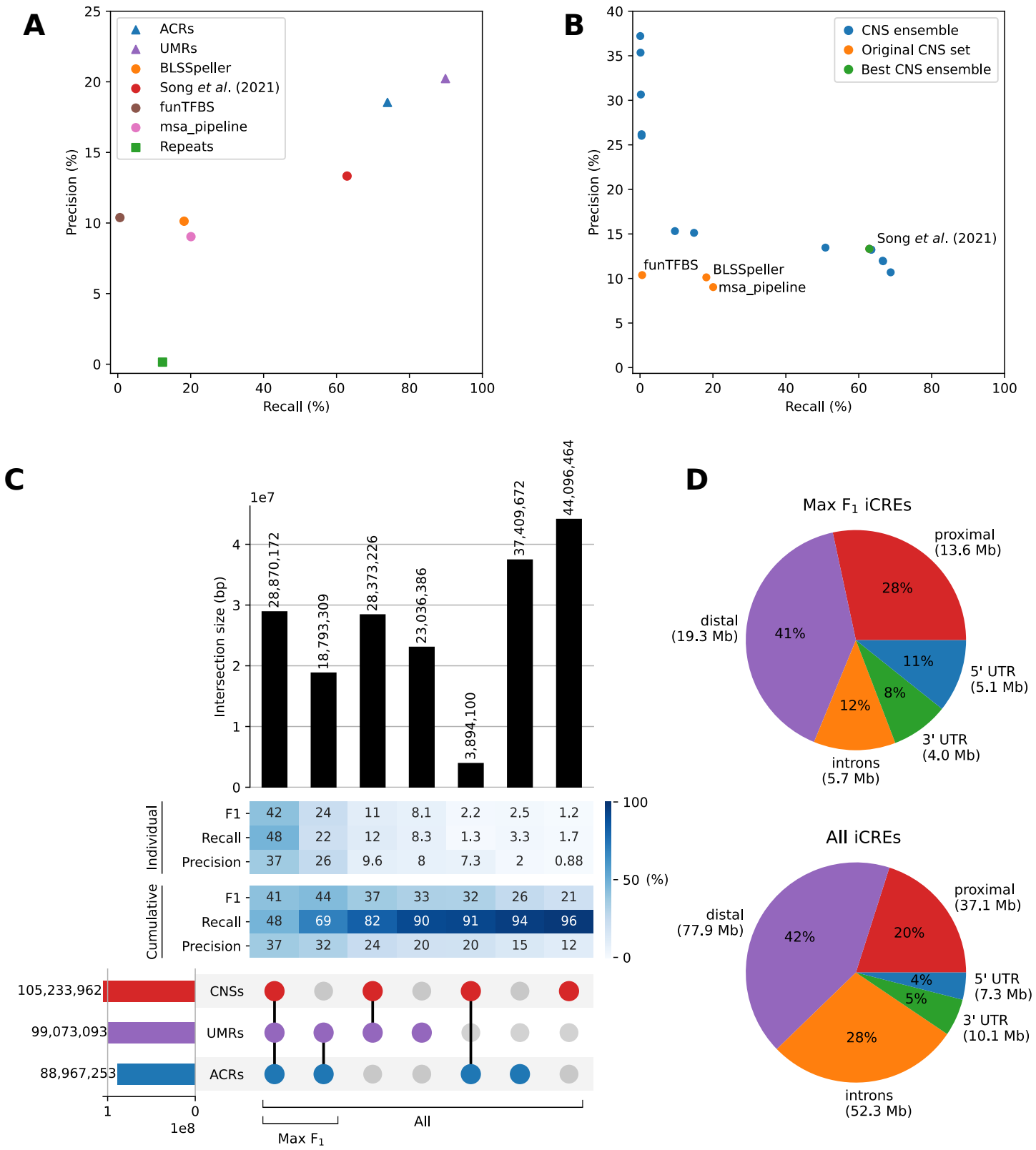

Figure 1. Benchmark, complementarity assessment, and integration of CNS detection and experimental CRE-profiling methods. (A) Precision-recall plot of individual CNS detection and experimental CRE-profiling methods, using a ChIP-seq dataset as gold standard. CNS detection methods are plotted as circles, experimental CRE-profiling methods as triangles, and control datasets as squares. (B) Precision-recall plot (using ChIP-seq as gold standard) of individual CNS detection methods and different ensembles that were constructed by combining the intersections of these methods. The ensemble with the best F1 is shown in green. (C) UpSet plot of the ACRs, UMRs and CNSs, with performance metrics of the subsets. The UpSet plot shows the sizes of the full datasets on the left and the sizes of the mutually exclusive subsets on the top. Precision, recall and F1 are calculated using a ChIP-seq gold standard for each of the individual subsets and are shown labeled as "individual". The UpSet plot is sorted on the precision of the subsets, so that the subset with the most correct predictions is shown on the left. The same performance metrics are also calculated cumulatively (labeled as "cumulative"), starting from the subset with the highest precision (most left) and progressively adding individual subsets with the next best precision, until all regions are added together (most right). The combined subsets that result in the "max F1 iCREs" and "all iCREs" are labeled "max F1" and "all" respectively. (D) The genomic feature distribution of the "max F1 iCREs" and "all iCREs". The total coverage in each genomic feature is given in megabases (Mb).

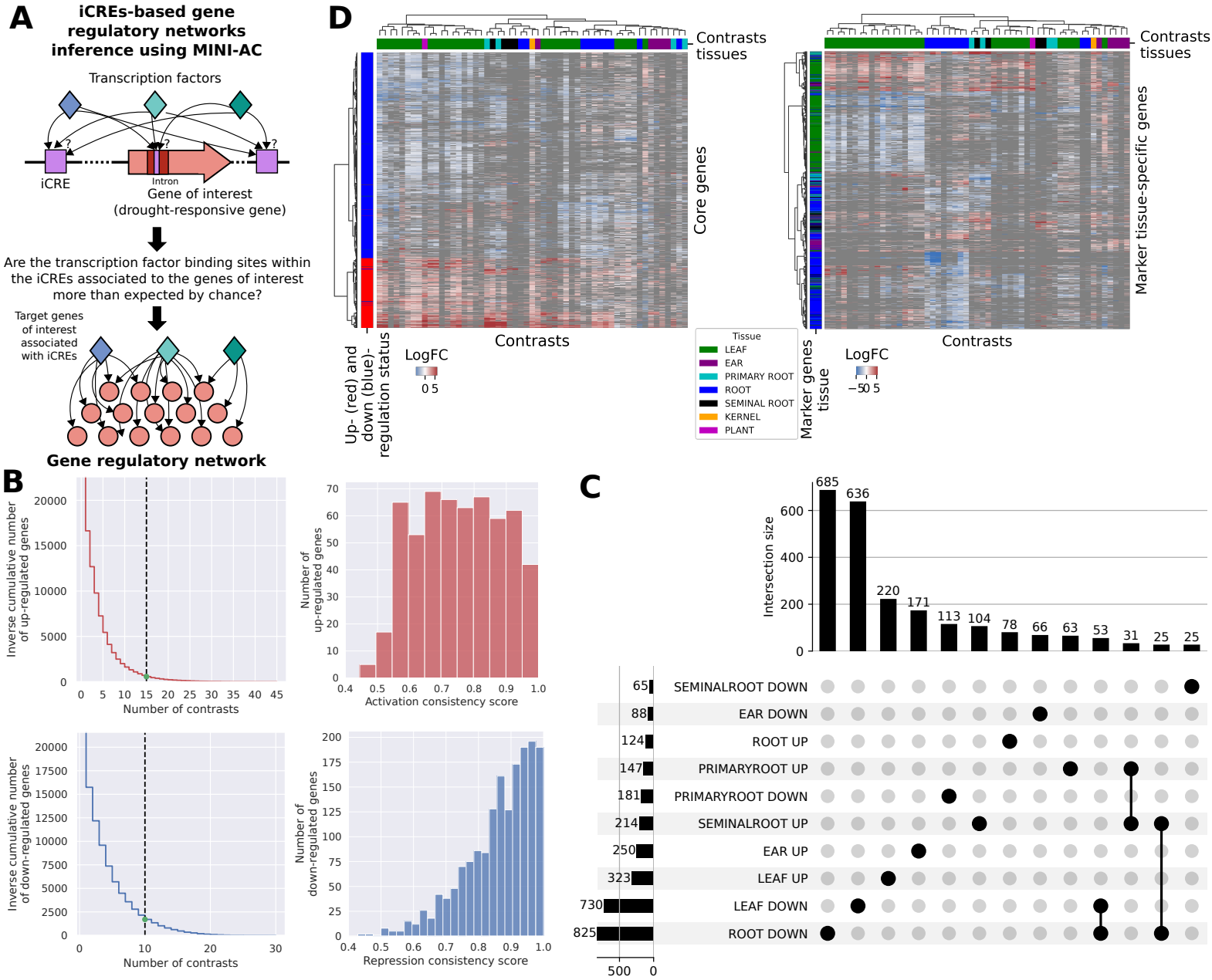

Figure 2. iCREs-based gene regulatory network (GRN) prediction and selection of drought-responsive gene sets. (A) Overview of the iCREs-based GRN inference framework, where the iCREs associated to a set of functionally-related or co-expressed genes are submitted to MINI-AC to perform motif enrichment and GRN inference. (B) On the left, there is the inverse cumulative distribution of the number of up- and down-regulated genes (y-axis) that are common in a specific number of contrasts (x-axis). On the right, there is the activation and repression consistency score distribution of the "core" drought-responsive genes. (C) UpSet plot showing the number of common and shared drought-responsive genes per tissue, sorted by the size of the intersection set. For visualization purposes, only intersection sets with 25 or more genes are shown. (D) Clustered heatmaps showing the log-transformed fold change of gene expression in the "core" (left) and "marker tissue-specific" (right) drought-responsive gene sets.

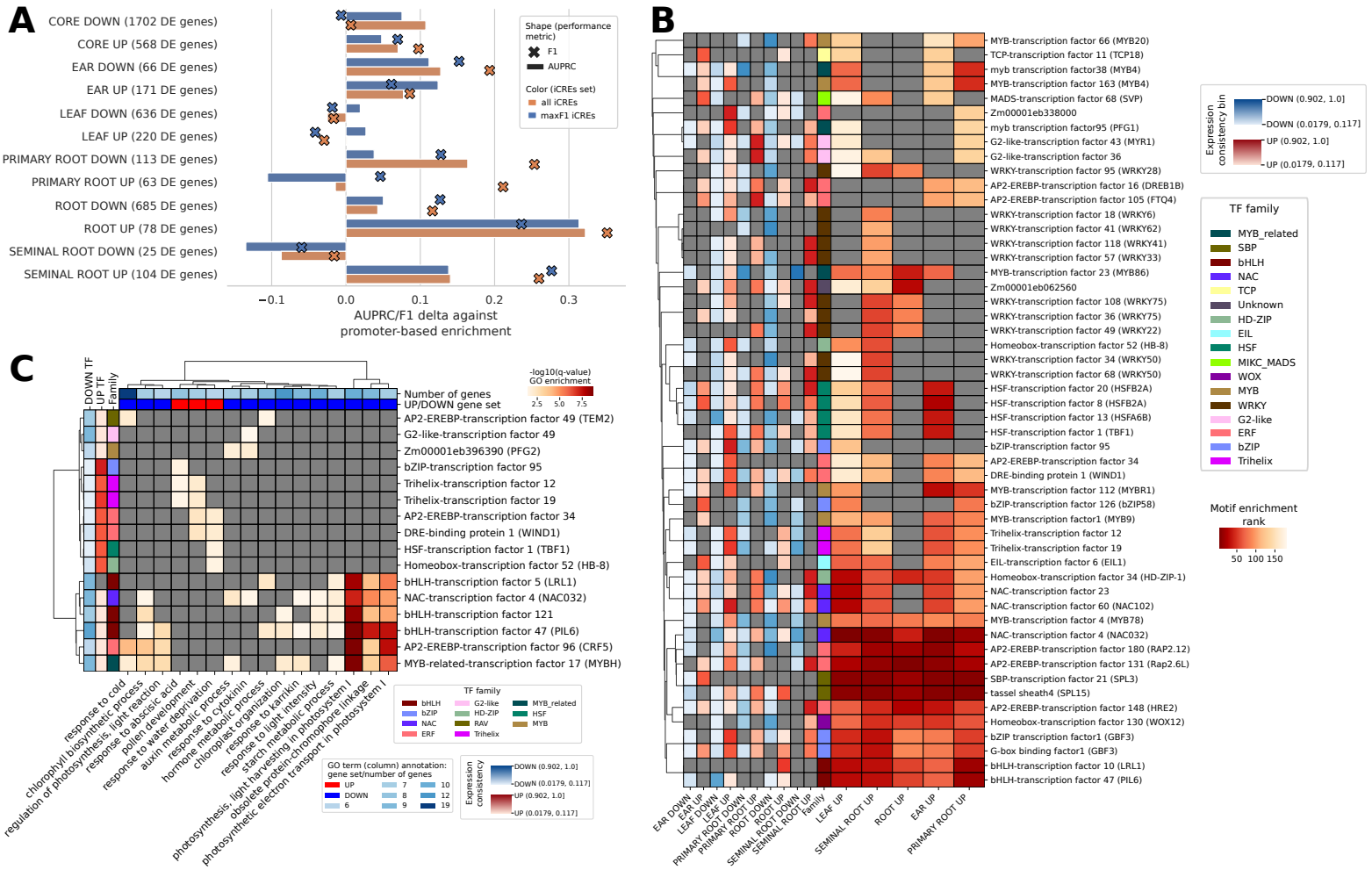

Figure 3. Benchmark and analysis of iCREs-based and drought-responsive gene regulatory networks (GRNs). (A) Comparison of the performance of the "promoter-based" and "iCREs-based" GRN inference frameworks, using two different iCREs sets (maxF1 and all). The two performance metrics (F1 and area under the precision recall curve) measure how well each method predicts differentially expressed (DE) regulators given a set of tissue-specific drought-responsive genes (y-axis). (B) Clustered heatmap showing the motif enrichment rank of different DE transcription factors (TFs) (y-axis) given sets of iCREs that are associated with tissue-specific (x-axis) up-regulated genes in response to drought. The row annotations correspond to the activation and repression consistency score bins and the TF family. For visualization purposes, only TFs with an expression consistency score higher than 0.513 are shown. (C) Clustered heatmap displaying the -log10(q-value) of gene ontology (GO; x-axis) enrichment among the predicted target genes of up- and down-regulated TFs in response to drought in leaf (y-axis). The row annotations indicate the activation and repression consistency score of the TFs, and the TF family. The column annotation corresponds to the gene set (up- or down-regulated marker genes in response to drought in leaf) that yielded the indicated GO enrichment, and the number of target genes in that GO category. For visualization purposes, the enriched GO terms with 5 or fewer target genes were discarded, as well as TFs with 3 or less GO terms enriched.

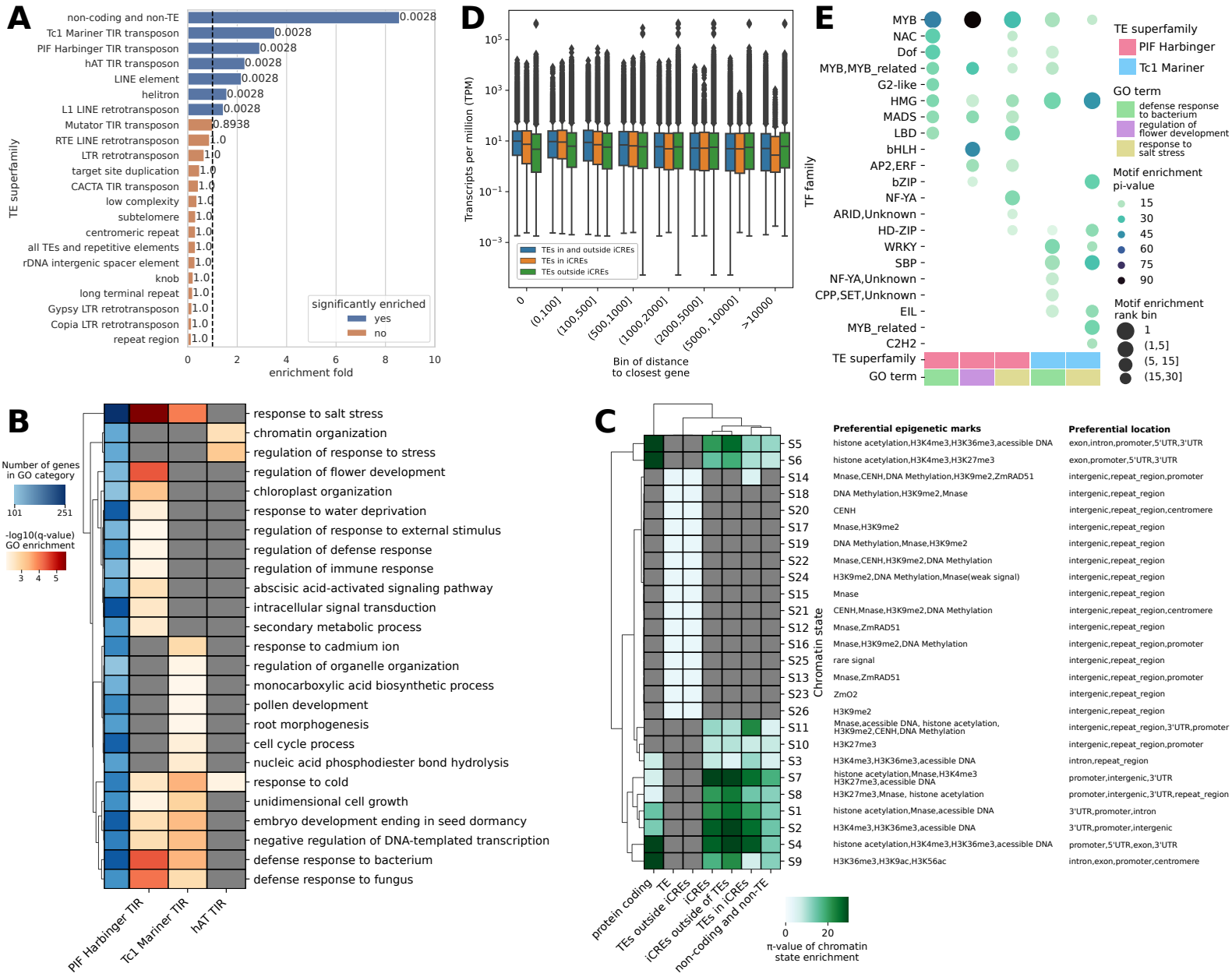

Figure 4. Analysis of the regulatory role of transposable elements (TEs) in maize using iCREs. (A) Bar plot that shows the enrichment fold (x-axis) and q-value of TE superfamilies and repetitive element types (y-axis) in the iCREs. (B) Clustered heatmap displaying the -log10(q-value) of gene ontology (GO; y-axis) enrichment among the genes associated with TEs within iCREs of three tandem inverted repeats (TIR) superfamilies (x-axis). The row annotations indicate the number of TE-associated genes in the corresponding GO category. The heatmap was filtered for TF-GO pairs with a q-value < 0.0008 with at least 100 TE-associated genes within that GO category. (C) Chromatin states (CS; y-axis) enrichment of different sets of genomic regions (x-axis), including all TEs and repetitive elements, iCREs, and TEs within and outside iCREs. The tables adjacent to the x-axis indicate the chromatin marks and genomic locations characteristic of each CS. (D) Box plot displaying the expression level distribution in transcripts per million (y-axis) of the genes associated with TEs within iCREs, outside iCREs, and both, binned per distance to the corresponding genes. (E) Bubble plot showing the motif enrichment rank and π-value (−log10(p-value) × enrichment fold) of different transcription factor (TF) families on the TEs within iCREs that are associated to genes annotated with “defense response to bacterium”, “regulation of flower development”, and “response to salt stress”, and that belong to the superfamilies Tc1/Mariner and PIF/Harbinger. Only the families with at least one motif with motif enrichment rank of 30 and less are shown.