Title: The genetic architecture of cell-type-specific *cis*-regulation

Authors: Alexandre P. Marand[1†*], Luguang Jiang[2], Fabio Gomez-Cano[2], Mark A.A. Minow[1], Xuan Zhang[1], John P. Mendieta[1], Ziliang Luo[1], Sohyun Bang[3], Haidong Yan[1‡], Cullan Meyer[1], Luca Schlegel[4], Frank Johannes[4], Robert J. Schmitz[1*]

Affiliations:

[1]Department of Genetics, University of Georgia, Athens, GA 30602, USA
[2]Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, MI, USA
[3]Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA
[4]Plant Epigenomics, TUM School of Life Sciences Weihenstephan, Technical University of Munich, Munich, Germany

Present Address:

[†]Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, MI, USA
[‡]College of Grassland Science and Technology, Sichuan Agricultural University, Chengdu, Sichuan 611130, China

***Corresponding authors.** Email: amarand@umich.edu and schmitz@uga.edu

---

Abstract:

Gene expression and complex phenotypes are determined by the activity of *cis*-regulatory elements. However, an understanding of how extant genetic variants affect *cis*-regulatory activity remains limited. Here, we investigated the consequences of *cis*-regulatory diversity using single-cell genomics of >0.7 million nuclei across 172 maize inbreds. Our analyses pinpointed *cis*-regulatory elements distinct to domesticated maize and how transposons rewired the regulatory landscape. We found widespread chromatin accessibility variation associated with >4.6 million genetic variants with largely cell-type-specific effects. Variants in TEOSINTE BRANCHED1/CYCLOIDEA/PROLIFERATING CELL FACTOR binding sites were the most prevalent determinants of chromatin accessibility. Finally, integration of genetic variants associated with chromatin accessibility, organismal trait variation, and population differentiation revealed how local adaptation has rewired regulatory networks in unique cellular context to alter maize flowering phenotypes.

Main Text:

Genome-wide association studies (GWASs) have identified thousands of genetic variants associated with agricultural traits. However, due to linkage disequilibrium (LD), disentangling large numbers of linked genetic variants to derive causal effects remains challenging even in species with low LD, such as *Zea mays* (maize). Maize trait-associated variants are mostly non-genic, and lie ~16.4 kb from the nearest gene, complicating efforts to infer the molecular mechanisms underlying phenotypic diversity (*1*). Thus, complementary approaches are required to augment GWAS results and pinpoint causal variants.

Variation in the patterns and magnitude of gene expression is a significant contributor to quantitative phenotypes. *Cis*-regulatory elements (CREs) orchestrate the spatiotemporal patterning and kinetics of gene expression, and have furthered understanding of the molecular determinants behind phenotypic manifestation (*2-4*). Indeed, studies have illustrated that CRE sequence evolution leads to morphological innovation and species divergence (*5*, *6*). Although some CRE variants have been associated with phenotypic changes (*7-12*), the mechanistic links between CRE variation and phenotypic diversity remain unclear. Moreover, different cell types are hallmarked by distinct CRE usage that is licensed by differential chromatin accessibility. Thus, defining the cellular contexts in which GWAS CRE variants act is a powerful avenue to understand polygenic trait manifestation (*13*).

Compared to humans (r^2=0.2 at 50-kb) (*14*) and mice (r^2=0.2 at 200-kb) (*15*), LD in maize is 10-40 times lower (r^2=0.2 at 5-kb) (*16*), presenting a unique opportunity to uncover the causal variants associated with molecular phenotypes. Taking advantage of this experimental system, we performed single-nuclei chromatin accessibility profiling in a diverse panel of 172 geographically distributed maize inbreds. We established the transcription factor (TF) binding sites contributing to accessible chromatin region (ACR) sequence conservation within maize and identified transposon-derived regulatory elements that shaped the regulatory landscape of domesticated maize. Leveraging the quantitative nature of chromatin accessibility, we revealed >4.6 M single nucleotide variants (SNVs) associated with

chromatin accessibility variability (chromatin accessibility quantitative trait loci, caQTL) at both the bulk tissue and cell-type scales. Comprehensive investigation of this dataset uncovered genetic variants within TF footprints for TEOSINTE BRANCHED1/CYCLOIDEA/PROLIFERATING CELL FACTOR (TCP) as significant determinants of chromatin accessibility and putative chromatin interactions. Finally, by integrating cell-type-resolved caQTL with GWAS hits, signatures of population differentiation, and TF binding sites, we shed light on the regulatory mechanisms that enabled tropical maize to adapt to temperate climates and the cellular contexts underlying complex trait variation.

Widespread intraspecific *cis*-regulatory variation

To establish the genetic determinants impacting *cis*-regulatory function at single-cell resolution in maize, we profiled chromatin accessibility from ~1.37 million seedling nuclei across 172 distinct maize inbred lines using single-cell Assay for Transposase Accessible Chromatin sequencing (scATAC-seq) with a pooling, replication, and *in silico* genotyping framework (**Fig. 1A**-**1C**, **fig. S1**; **table S1**, **S2**). Following strict thresholding for genotype calls (**fig. S2A**), we retained 682,794 high-quality nuclei (average of 3,970 nuclei per inbred) and identified a total of 108,843 ACRs (**Fig. 1C**; **table S3**). Variant calls and global chromatin accessibility profiles were consistent with prior genotyping and among replicated genotypes, respectively (**fig. S2B**-**S2E**).

We adopted a reference mapping approach to ensure robust cell-type identification (**fig. S3A**-**S3C**). Specifically, a reference set of high-quality nuclei (N=82,283; Materials and Methods) from all 172 genotypes were clustered into 52 distinct groups with similar genome-wide chromatin accessibility profiles providing a reference embedding to project the remaining nuclei (**Fig. 1D**). To assign cell identity to each cluster, we estimated gene chromatin accessibility scores for all nuclei and assessed chromatin accessibility dynamics for known marker genes via an iterative strategy composed of (*i*) automated annotation based on cell-type and cell cycle-stage enrichment scores, (*ii*) focused cell-type annotation refinement by evaluating differentially accessible chromatin of known cell type and spatial marker genes, and (*iii*) visual inspection of marker gene enrichment (**Fig. 1E**, **fig. S3D**; **table S4**). The

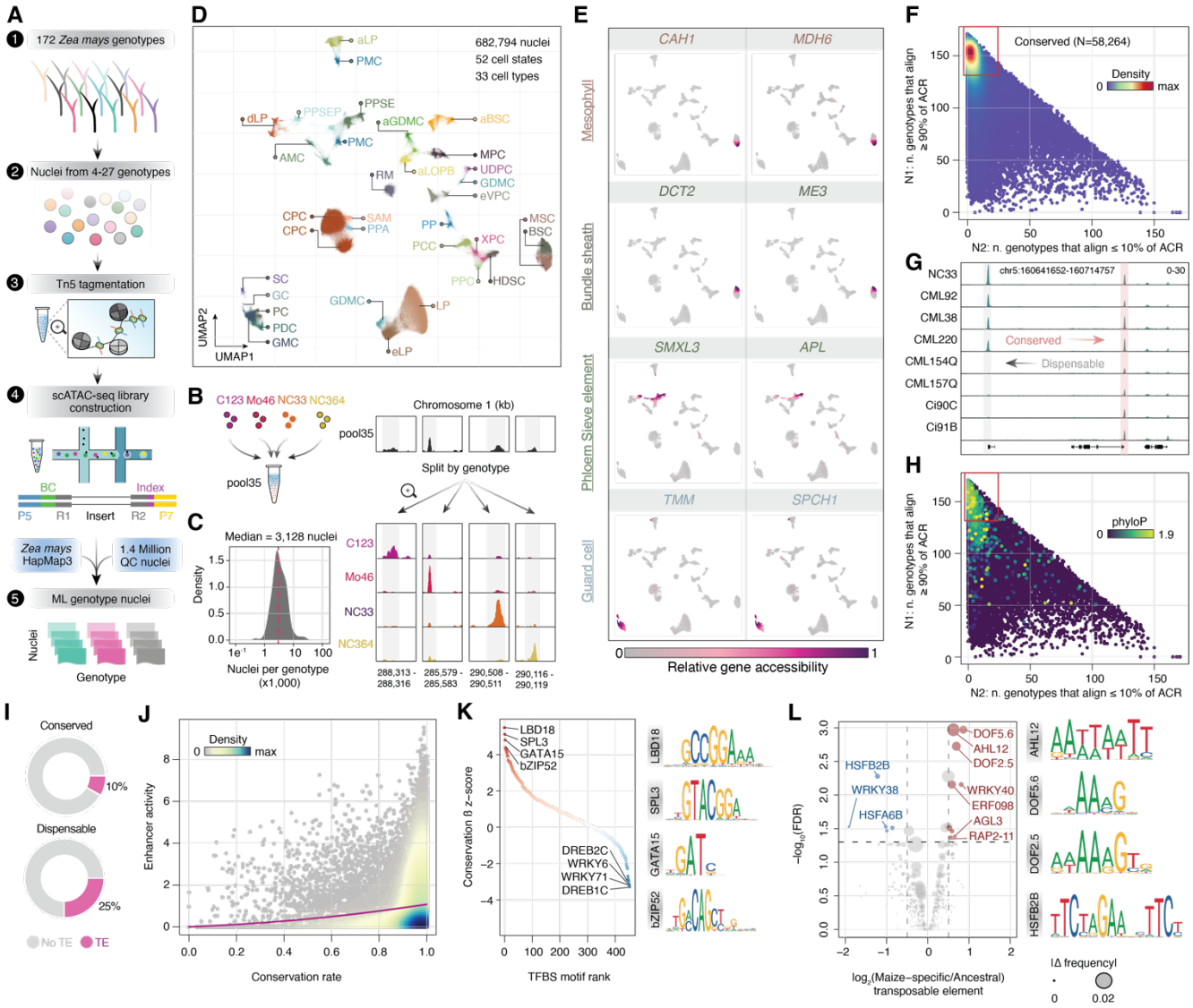

Fig. 1. Widespread *cis*-regulatory sequence variation. (**A**) Experimental and computational workflow. (**B**) Pooling strategy and genotype deconvolution. (**C**) Distribution of nuclei counts per genotype. (**D**) UMAP visualization of chromatin accessibility variance among nuclei from diverse maize genotypes. (**E**) Visualization of chromatin accessibility variance for eight major cell-type-specific marker genes. (**F**) Scatter plot illustrating the number of genotypes with greater than 90% coverage from whole-genome resequencing of an ACR (y-axis) versus the number of genotypes with less than 10% coverage of an ACR (x-axis). Points are colored by density. (**G**) Genome browser examples of conserved and dispensable ACRs identified from panel F. (**H**) Same as panel F, except ACRs are colored by weighted average phyloP scores among grass species. (**I**) Pie charts comparing the proportion of ACRs overlapping transposable elements from the conserved and dispensable classifications. (**J**) Scatter plot of ACR conservation rate (x-axis) and enhancer activity (y-axis) as estimated by STARR-seq. Points are colored by density. (**K**) Ranked scatter plot of permutation normalized TFBS effects towards ACR conservation rate adjacent to exemplary TF motifs. (**L**) Bubble plot of the $\log_2$ transformation of the ratio between the fraction of maize-specific TE-ACRs and the fraction of ancestrally conserved TE-ACRs containing a given TF motif (x-axis) against the $\log_{10}$ transformed FDR. The size of the bubble reflects the absolute difference in TE-ACR frequency for a given TF motif. Exemplary TF motifs are illustrated on the right.

final cell-type (N=33), cell cycle-stage, and marker-based spatial labels (hereafter referred to as cell state) were uniformly distributed among genotypes and supported through an integration with 30,305 nuclei generated from single-nucleus RNA sequencing (snRNA-seq) (**fig. S4**, **fig. S5**; **table S5**).

Genome-wide documentation of intraspecific *cis*-regulatory variation is limited (17). To understand the relationship between genetic variation and ACR conservation, we evaluated sequence conservation of 82,098 non-coding ACRs in the context of the B73 reference genome and found 58,264 (71%) ACR sequences conserved among genotypes (**Fig. 1F-1G**). The remaining dispensable ACRs were (*i*) associated with significantly lower evolutionary conservation rates (phyloP; 5.74-fold depletion, Wilcoxon rank sum test, WRST; $P<2.2e^{-16}$; **Fig. 1H**; **table S3**), (*ii*) overrepresented in intergenic and gene-proximal regions (**Fig. S6A**), and (*iii*) overlapped significantly with certain repetitive sequences (Fisher's exact test, FET; $P<2.2e^{-16}$; **Fig. 1I**, **fig. S6B**). Showcasing the functional importance of conserved ACRs, we found a significant positive relationship between conservation rate and STARR-seq enhancer activity (Likelihood ratio test, LRT; $P<2.2e^{-16}$; **Fig. 1J**) (18). We hypothesized that constraint on the activities of specific transcription factor binding sites (TFBSs) could underlie ACR conservation. Indeed, a Monte Carlo permutation ridge regression test integrating resequencing information associated the occurrence of 64 TFBS motifs with elevated ACR conservation (FDR<0.05; **Fig. 1K**; **table S6**). Most of these highly conserved TFBS motifs recruit critical developmental regulators, including TFs from *LATERAL ORGAN BOUNDARY DOMAIN* (*LBD*) (19), *SQUAMOSA PROMOTER BINDING PROTEIN-LIKE* (*SPL*) (20), and *GATA* gene families (21). TFBS motifs from the environmentally responsive *WRKY* and *DEHYDRATION RESPONSE ELEMENT BINDING* (*DREB*) gene families exhibited strong negative association with conservation, consistent with stimuli-response CREs contributing to divergence and local adaptation (**Fig. 1K**) (22, 23).

By isolating ACRs (n=9,174) lacking sequence alignments to the wild progenitor, *Z. mays* ssp. *parviglumis* (teosinte; **table S3**), we identified ACRs unique to domesticated maize that should contain CRE innovations that underlie the transition to cultivated maize. Consistent with a recent origin, domesticated maize-specific ACRs were enriched for dispensability (FET, $P<2.2e^{-16}$), lower enhancer activity (WRST, $P<2.2e^{-16}$), and increased overlap with repetitive elements (FET, $P<2.2e^{-16}$; **fig. S6C-S6E**). Motivated by historical evidence of transposable element-derived (TE) CRE sequences in the maize genome (9, 12, 24, 25), we asked if certain TE families played a role in re-wiring TF targets by

focusing on a set of 1,587 ACRs fixed in domesticated maize and absent in teosinte. We found a significant enrichment of *hAT* and *PIF/Harbinger* DNA TE families within 483 ACRs post-dating teosinte divergence (Chi-square test, CST; FDR<0.05; **fig. S6F**). The TEs overlapping domesticated maize-specific ACRs were enriched for TFBSs corresponding to DNA-BINDING WITH ONE FINGER (DOF) (26), ETHYLENE RESPONSE FACTOR (ERF) (27), and SEPALLATA (SEP) (28) TFs, which regulate vasculature and floral development (**Fig. 1L**). *PIF/Harbinger* insertions are known to affect the floral transition by modulating vascular expression of *CONSTANS*, *CONSTANS-LIKE*, and *TIMING OF CAB1* (*CCT*) clock genes (7, 12), consistent with our genome-wide results. These results highlight that co-option of *hAT* and *PIF/Harbinger* CREs may have played a widespread role in rewiring the regulatory landscape of domesticated maize.

Genetic determinants of chromatin accessibility

Single-cell data allows investigation into the effects of *cis*-regulatory variation at cell-type resolution. Focusing on the 57,984 ACRs with significant genotype or cell-state effects (**fig. S7A**; ANOVA, FDR<0.05) and cell states with sufficient power (N=20), we performed caQTL mapping by modeling normalized chromatin accessibility as a function of single nucleotide variant (SNV) allelic dosage and a suite of latent and measured covariates (Materials and Methods). We uncovered 4.6 million *cis*-caQTL associated with 23,959 and 23,123 ACRs (hereafter termed caQTL-ACRs; dynamic beta-adjusted FDR<0.1–0.01) at the bulk-tissue and cell-state-resolved scales, respectively (**Fig. 2A-2C**, **fig. S7B-S7E**). Several caQTL-ACRs represented known CRE variants with trait effects, including those regulating *ZmCCT9* (7), *Vgt1* (8), *gt1* (11), and *Upright Plant Architecture 2* (10) (**fig. S8A**). Additionally, bulk caQTL overlapped significantly (26%; CST, $P<2.2e^{-16}$) with previously identified expression QTL (eQTL) (29) with ~85% concordance in effect directions (**Fig. 2D-2E**, **fig. S8B-S8F**). Most alternate alleles (76%) were associated with both reduced chromatin accessibility and transcription, indicating that a majority of alternate allele caQTL disrupt transcription-activating CREs (**Fig. 2E**). However, we also find matched caQTL-eQTL variants associated with increased chromatin accessibility and transcription (class ii; **fig. S8D**). These variants were enriched in Polycomb Response Elements (PREs)

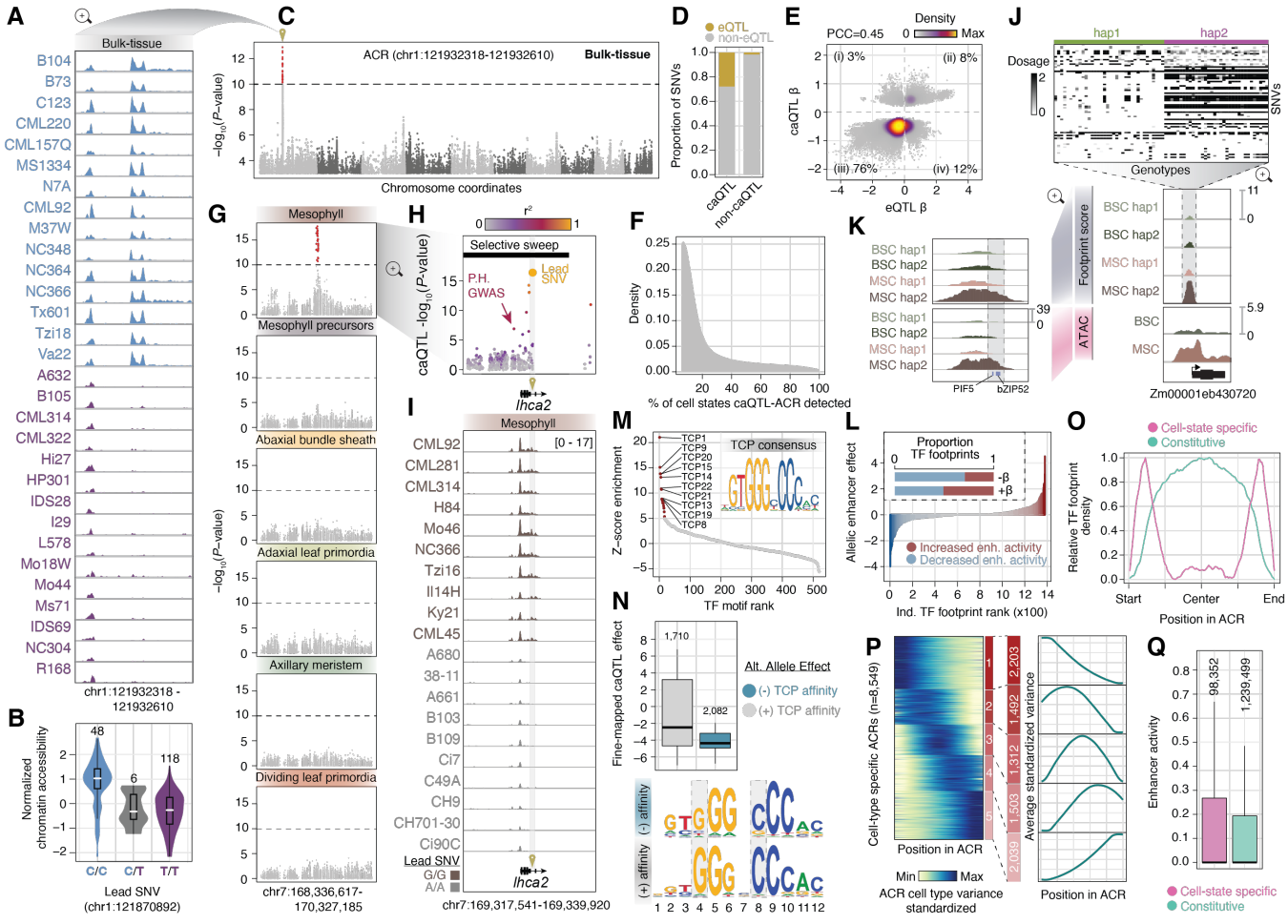

Fig. 2. Genetic determinants of chromatin accessibility variation. (**A**) Genome browser view of a caQTL-ACR in 30 exemplary bulk-tissue aggregates colored by the lead SNV genotype in panel B. (**B**) Violin plot of the distribution of normalized chromatin accessibility for three caQTL genotypes associated with the ACR in panel A. (**C**) Manhattan plot of SNV significance by genomic position for a single ACR phenotype in pseudobulk-tissue aggregates. (**D**) Coincidence of caQTL and eQTL between the closest ACR-gene pairs from bulk-tissue association mapping. (**E**) Comparison of caQTL and eQTL effect sizes and directions for SNVs from matched ACR-gene pairs. The percentage of variants are denoted in each quadrant. (**F**) Distribution of caQTL-ACRs detection across cell states. (**G**) Manhattan plot for a cell-state-specific caQTL. (**H**) Variant association plot versus genomic position for the ACR shown in panel I. Points are colored by linkage disequilibrium (r^2) relative to the lead variant. A previous GWAS hit for plant height is shown by the red arrow. The location of a selective sweep is shown as a black bar. (**I**) Genome browser view of 20 exemplary genotypes at the *lhca2* locus aligned with the genomic coordinates of panel H. Tn5 integrations are scaled per million. (**J**) Example of haplotype clustering and TF footprinting in bundle sheath and mesophyll cells. (**K**) Zoom-in on allele-specific footprint scores (top) and chromatin accessibility (bottom). (**L**) Main: Allelic difference in enhancer activity for 1,379 TF footprints. Inset: Proportion of TF footprints with decreased or increased enhancer activity grouped by caQTL effect direction. (**M**) TF footprint motifs ranked by enrichment with fine-mapped caQTL. Red dots indicate FDR < $1e^{-5}$. (**N**) Top: Distribution minus outliers of fine-mapped caQTL effects for SNVs increasing and decreasing TCP binding affinity. Bottom: Motif logos from variants predicted to increased and decrease TCP binding affinity. (**O**) Distribution of cell-state-specific and constitutive TF footprints across ACRs. (**P**) Heatmap illustrating standardized ACR variance across cell types clustered by k-means. The average cluster standardized variances are shown to the right. (**Q**) Distribution minus outliers of enhancer activity for cell-state specific and constitutive TF footprints.

recognized by the BASIC PENTACYSTEINE (BPC) family of TFs (BPC5, BPC6, and BPC1) relative to background (*30*, *31*), suggesting that variants inhibiting BPC-based Polycomb Repressive Complex 2 targeting arise frequently and increase local chromatin accessibility and gene expression (FET; $P<7.1e^{-5}$; **fig. S9A**).

We posited that bulk tissue cellular heterogeneity masks caQTL from rare cellular contexts. Indeed, 44% of caQTL-ACRs identified by cell-state-resolved caQTL mapping were not found in bulk tissue and were associated with greater ACR cell-state specificity (WRST; $P<3.5e^{-112}$; **fig. S9B**-**S9C**). Similarly, 63% of cell-state-resolved caQTL-ACRs were specific to a single cell state and often overlapped with variants associated with complex traits (**Fig. 2F**-**2G**). For example, in the last intron of *light harvesting complex a2,* we identified a mesophyll-specific caQTL-ACR missed in bulk tissue which was linked (r^2=0.6) with a GWAS hit for plant height and embedded within a selective sweep in temperate non-stiff stalk germplasm (**Fig. 2H**-**2I**) (*32*). As expected, caQTL-ACRs identified in multiple cell states and abundant cell types were more likely to be identified in bulk tissue and have 100% concordant effect directions, whereas cell-state-specific caQTL frequently (11%) exhibited opposing effect directions in other cell types (**fig. S9D**-**S9H**).

caQTL perturb TFBSs

We reasoned that caQTL variants within TFBSs were the most likely to be causal. Focusing on the cell-state-resolved caQTL-ACRs, we fine-mapped 22,053 putatively causal caQTL (95% credible sets) across 4,815 caQTL-ACRs with an average of 3,054 putative causal variants per cell state (**fig S10A**; **table S7**). Next, we identified differential TF footprints between the two most frequent haplotypes within each cellular context, reflecting sites with differential TF occupancy between alleles (**Fig. 2J**). This analysis revealed an average ~25,000 allele-specific TF footprints within each cell state, including examples of allele- and cell-state-specific chromatin accessibility over TF footprints (**Fig. 2K**, **fig. S10A**). Indicative of regulatory activity, TF footprints had higher STARR-seq enhancer activity compared to nearby controls (50-bp shift; WRST; $P<6.9e^{-101}$; **fig. S10B**). Nearly half (47%) of fine-mapped caQTL were in TF footprints, a 50-fold enrichment compared to randomized control regions (Monte Carlo permutation, MCP; $P<1e^{-4}$; **fig. S10C**). To validate the effects of the fine-mapped variants within TF footprints, we used STARR-seq data to quantify enhancer activity from both reference and alternate alleles. Fine-mapped TF footprint variants had significantly greater absolute effect sizes on

STARR-seq enhancer activity compared to non-caQTL variants (WRST; $P<2.6e^{-73}$; **fig. S10D**). Moreover, TF footprint alleles associated with negative effect caQTL had a greater frequency of decreased enhancer activity (CST; $P<9.8e^{-8}$; **Fig. 2L**). Cell-state resolved analysis of TF footprint enrichment was consistent with known TF usage among cells states; footprint enrichment was observed for REVEILLE1 (RVE1) in mesophyll (*33*), DNA-BINDING WITH ONE FINGER (DOF) in vasculature (*34*), KANADI (KAN) in abaxial bundle sheath (*35*), and E2F in dividing leaf primordia cells (**fig. S10E-S10G**) (*36*). We observed coincidence of cell-state specificity among fine-mapped caQTL, differential TF footprints, and chromatin accessibility, suggesting that cell-state-specific TFBS variants precisely alter chromatin accessibility in the cognate cell state (MCP, $P<1e^{-4}$; **fig. S10H**). We then assessed the impact of fine-mapped SNVs on TF-binding affinity. Fine-mapped caQTL were associated with significantly greater absolute allele effects compared to non-caQTL SNVs (MCP; $P<1e^{-4}$), suggesting that putatively causal caQTL affect chromatin accessibility by influencing TF binding affinity (**fig. S10I-S10J**).

We hypothesized that only select TFBS variants affect chromatin accessibility. To test this, we compared the enrichment of TF footprint classes between SNVs within fine-mapped caQTL and non-caQTL. Fine-mapped caQTL within TFBSs recognized by TCP family members had the strongest effects (MCP, FDR<$1e^{-5}$; **Fig. 2M**). Moreover, fine-mapped variants disrupting the TCP binding motif, particularly at the 4th and 8th base positions, lost chromatin accessibility (WRST; $P<5.6e^{-8}$; **Fig. 2N**). The transcriptional signatures of TCP TFs were largely cell-type-specific, suggesting that distinct TCP TFs promote chromatin accessibility in unique cellular contexts (**fig. S11**).

Cell-state-specific TF footprints tended to occur on ACR edges (**Fig. 2K**). Indeed, cell-state-specific TF footprints were biased towards ACR boundaries compared to TF footprints present across multiple cell states (**Fig. 2O**). To measure chromatin accessibility dynamism at ACR boundaries, we estimated, and k-means clustered, the cell-state variance in chromatin accessibility for each cell-state-specific ACR (N=8,549) along the length of ACR sequence. Approximately 86% of cell-state-specific ACRs were

most variable in boundary regions (clusters 1, 2, 4, 5; **Fig. 2P**). Quantifying cell-state-resolved TF footprint frequencies spatially within ACRs (**fig. S12A**) revealed cell-type-specific TFBS enrichment in ACR peripheries (**fig. S12B**). These boundary TFBSs had greater STARR-seq enhancer activity compared to constitutively accessible TFBSs (WRST; $P<1.2e^{-12}$; **Fig. 2Q**), indicating dynamic ACR boundaries unveil potent cell-type-specific CREs. Notably, consistent across cell states, the top 20 most enriched TF footprints within ACR centers all corresponded to TCP TFs, further supporting a role in promoting accessible chromatin (**fig. S12C**).

Genetic determinants of long-range chromatin interactions

More than 56% of bulk-scale caQTL were associated with more than one ACR (**fig. S13A**), suggesting a single caQTL might impact multiple *in nucleo* spatial interactions. We identified 29,281 long-range caQTL (29,293 ACR-ACR linkages), defined as a caQTL variant associated with chromatin accessibility changes at more than one linked ACR (**Fig. 3A**). Long-range caQTL had a bimodal distribution of ACR-ACR distances biased towards intergenic interactions with an average of 3.1 genetic interactions per ACR (**fig. S13B**-**S13C**). Moreover, long-range caQTL were enriched for fine-mapped variants (FET; $P<2.2e^{-16}$), co-accessible ACRs (MCP, $P<1e^{-4}$), predicted chromatin interactions based on a deep learning convolution neural network with a long short-term memory (MCP, $P<1e^{-4}$) (*37*), and HiC and HiChIP chromatin loops (MCP, $P<1e^{-4}$) relative to randomly permuted caQTL-ACRs links (**Fig. 3B**-**3C**; **fig. S13D**). For example, we identified a caQTL variant within the promoter of *abscisic acid insensitive 8* (*abi8*), an ABI-VP1 TF with a hypothesized negative regulatory role in leaf blade morphogenesis (*38*). The *abi8* promoter variant was associated with ACRs 10-kb and 50-kb downstream and coincident with H3K4me3- and H3K27me3-HiChIP chromatin loops (**Fig. 3D**). Additionally, long-range caQTL with fine-mapped variants showed stronger overlap with chromatin loops than those lacking fine-mapped caQTL (FET; $P<1.49e^{-6}$), suggesting fine-mapped caQTL frequently affect distal ACRs.

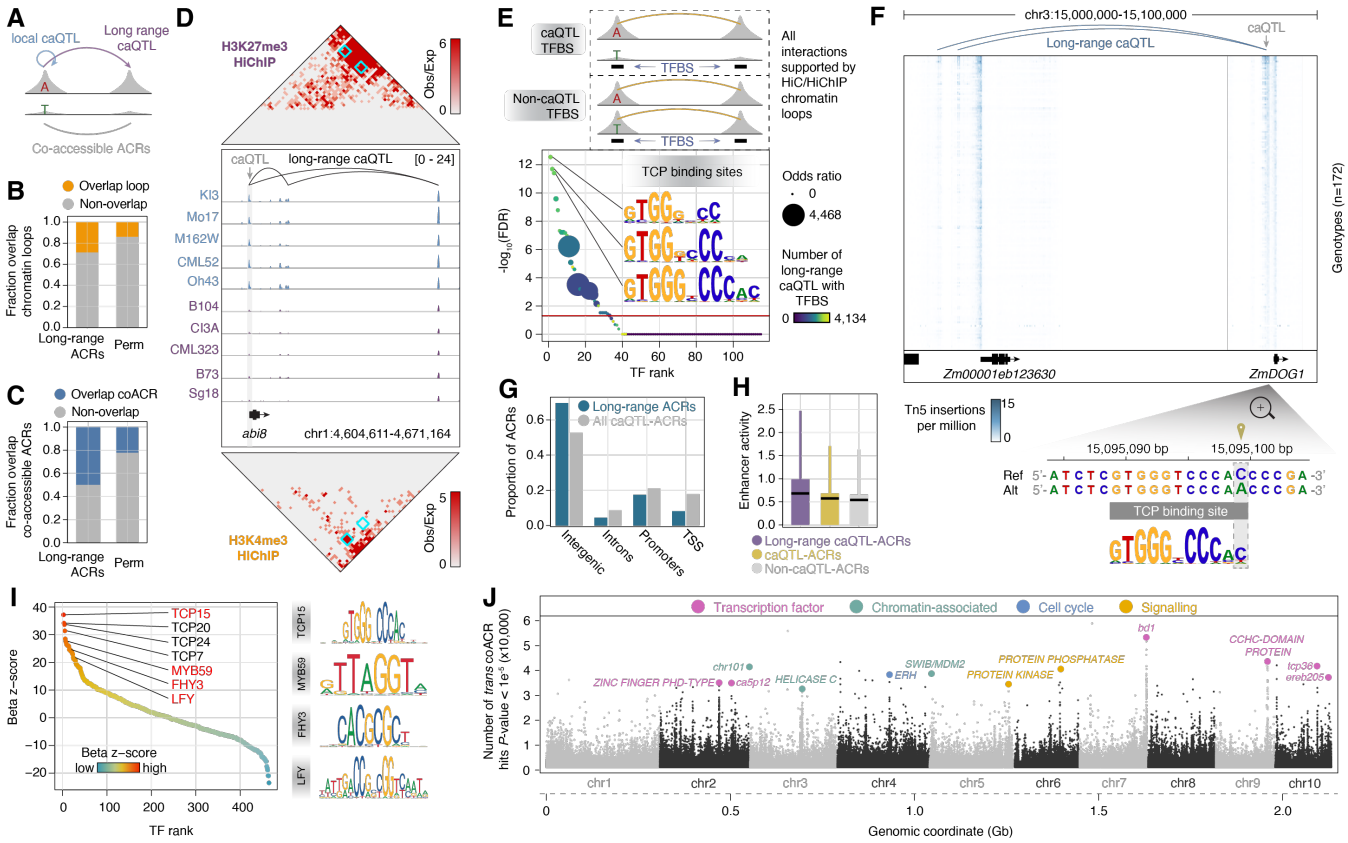

Fig. 3: Long-range caQTL perturb chromatin loops. (**A**) Schematic of local and long-range caQTL (top), and co-accessible ACRs (bottom). (**B**) Fraction of predicted (based on long-range caQTL genetic evidence) and permuted ACR-ACR interactions overlapping experimentally resolved chromatin loops. (**C**) Fraction of predicted and permuted ACR-ACR interactions overlapping with co-accessible ACRs. (**D**) Middle: Exemplary long-range caQTL interaction between the promoter of *abi8* and two ACRs downstream. Top: Normalized H3K27me3 HiChIP contact matrix. Bottom: Normalized H3K4me3 HiChIP contact matrix. Blue rectangles denote the interaction frequency between the promoter of *abi8* and the downstream ACRs. (**E**) Top, illustration of test sets (TFBS counts containing long-range caQTL or non-caQTL SNVs conditioned on HiC/HiChIP looping and reciprocal presence of TFBSs at both anchors). Bottom, bubble plot of TF enrichment ranks by -$\log_{10}$(FDR). Points are colored by the number of interactions and sized by the odds ratio from Fisher's exact tests. Red line indicates -$\log_{10}$(FDR = 0.05). Inset, TFBS logo of top three enriched motifs corresponding to 17 TCP TF family members. (**F**) Top, density heatmap of Tn5 insertion counts across genotypes (rows). Bottom, zoom-up of caQTL alleles overlapping a predicted TCP binding site. (**G**) Proportion of ACR in each genomic context split by long-range caQTL-ACRs with local caQTL and all caQTL-ACRs. (**H**) Distribution of enhancer activity across long-range caQTL-ACRs, all caQTL-ACRs, and non-caQTL-ACRs. Black bars indicate distribution means. (**I**) TF motifs ranked by beta coefficient Z-scores (relative to 100 permutations) from a generalized linear model of long-range caQTL-ACR interactive capacity. (**J**) Manhattan plot of the number of co-accessible ACRs associated (*P*-value < $1e^{-5}$) with a single SNV. Candidate genes are annotated by different colors.

To understand the regulatory basis of long-range caQTL, we compared TFBSs found in both ACRs of long-range interactions supported by HiC and HiChIP chromatin loops containing either fine-mapped caQTL or non-caQTL SNVs (**Fig. 3E**). Like variants that impact local chromatin accessibility, caQTL affecting long-range interactions were enriched for TCP family TFBSs (**Fig. 3E**). For example, a local caQTL with variation in a predicted TCP binding site in the promoter of the maize homolog of *DELAY OF GERMINATION 1* (*DOG1*) was associated with two distal co-accessible ACRs 88-kb and 90-kb upstream (**Fig. 3F**). Interestingly, long-range caQTL within intergenic ACRs were more prevalent

than other genomic contexts (CST; $P<1.7e^{-221}$), associated with a greater number of interactions (interactive capacity), and had stronger STARR-seq enhancer activity (WRST; $P<2.0e^{-28}$; **Fig. 3G-3H**, **fig. S13E**). To evaluate whether specific TFBSs relate to interactive capacity, we fit a generalized linear model of interactive capacity as a function of motif occurrences within long-range caQTL-ACRs and compared the model coefficients with 100 random permutations. Consistent with preceding analyses, TCP family TFs were strongly associated with increased interactive capacity (**Fig. 3I**). These results suggest that gene-distal enhancers containing TCP binding sites may be determinative of long-range chromatin looping, consistent with past studies connecting TCP TFs to plant chromatin interactions (*39*).

Identifying the *trans*-regulators driving chromatin interactions remains difficult in plants. To address this, we performed QTL mapping on co-accessible ACRs supported by HiC/HiChIP loops (N=802,504). This analysis revealed 12 genes associated with genome-wide co-accessibility (**Fig. 3J**), including six transcription factors, three chromatin-associated proteins, two phospho-signaling proteins, and a cell-cycle associated gene (**table S8**). Interestingly, we identified a haplotype containing *tcp-transcription factor 36* (*tcp36*) on chromosome 10 as one of the top candidates associated with variation in chromatin co-accessibility (**fig. S13F**). The coding sequence for *tcp36* contained several missense mutations and a premature stop codon, adding support for the association between TCP TFs and global chromatin accessibility variation.

Selection on regulatory variants shaped complex traits

To understand the relative contribution of regulatory variation towards organismal phenotypes and better define the molecular modules underlying complex traits, we performed chromatin accessibility genome-wide association (CAWA) and transcriptome-wide association mapping (TWA). Phenotypes clustered by physiological relationship and associations were both consistent between linked ACRs and genes and captured known gene-trait interactions (**Fig. 4A**; **table S9-S10**). For example, transcript abundances and regulatory region chromatin accessibilities for *LONELY GUY 7* (*ZmLOG7*), *CUP-*

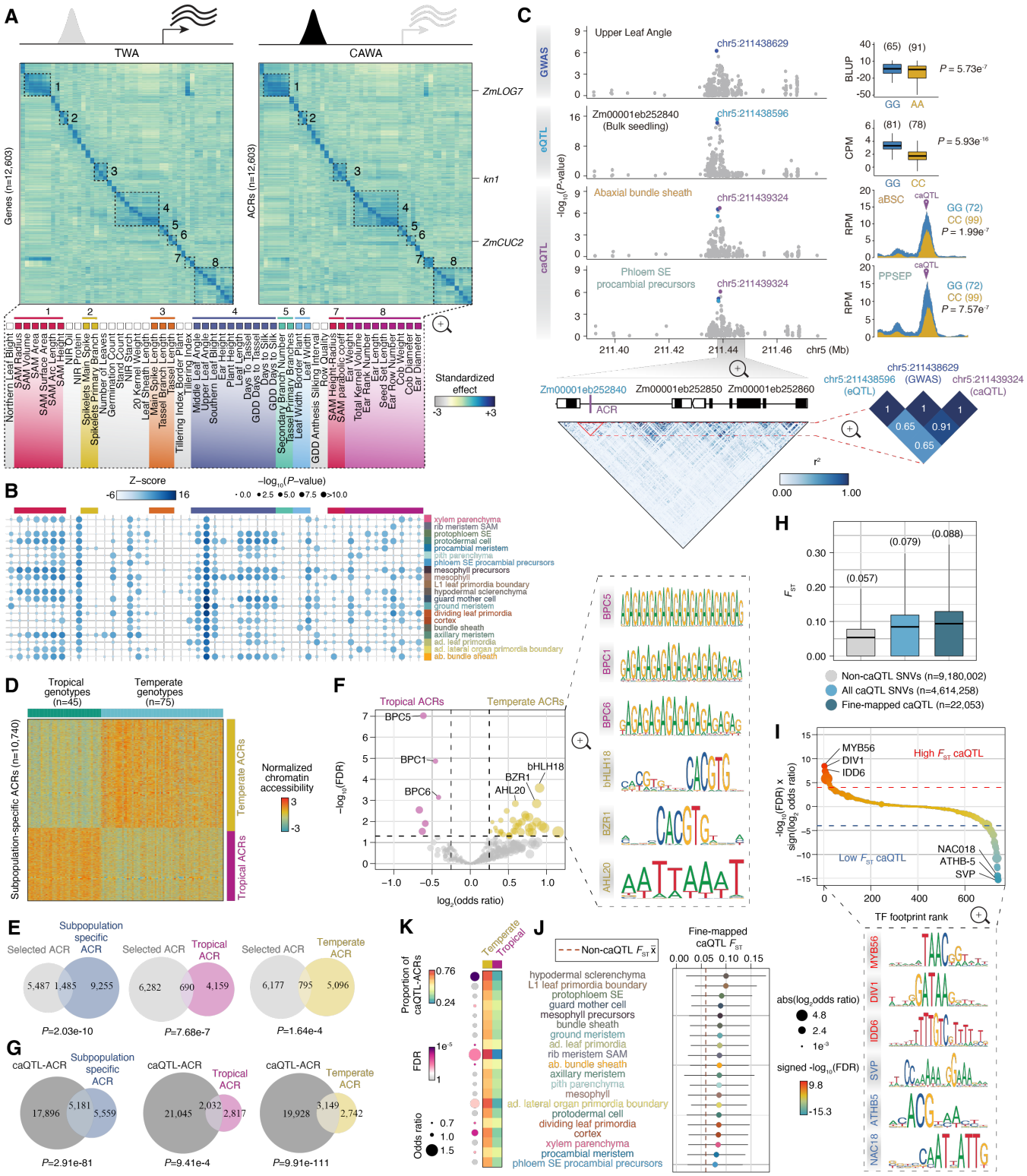

Fig. 4: Integration of caQTL with eQTL and GWAS of complex traits. (**A**) Heatmaps of standardized effects from CAWAS and TWAS for matched ACRs (left) and genes (right). Phenotype traits of the columns (identical order for both heatmaps) are shown below. (**B**) Permutation tests (x10,000) of percent variance explained by caQTL partitioned by cell state compared to non-caQTL SNVs. (**C**) Manhattan plots of upper leaf angle GWAS (top), expression QTL of bulk seedlings (middle), and caQTL for abaxial bundle sheath and phloem sieve element/procambial precursors (bottom). Trait distributions split by the lead variant genotypes are shown to the right. Estimates of linkage disequilibrium across variants within a focused 10-kb window are shown on the bottom as well as r^2 values among the three lead SNVs. (**D**) Heatmap of subpopulation-specific normalized chromatin accessibility (rows) across 120 tropical and temperate maize genotypes. (**E**) Venn diagram of ACRs within genomic regions under selection that overlap subpopulation-specific ACRs, tropical-specific ACRs, and temperate-specific ACRs. (**F**) Volcano

plot of differentially accessible motifs between tropical and temperate ACRs. (**G**) Venn diagram of caQTL-ACRs that overlap subpopulation-specific ACRs, tropical ACRs, and temperate ACRs. (**H**) Distribution of F_{ST} values of non-caQTL SNVs, all caQTL SNVs, and fine-mapped caQTL SNVs. (**I**) TF footprint ranked by the signed significance ($-\log_{10}$ FDR) for enrichment of high or low F_{ST} SNVs. (**J**) The distribution (standard deviation) and mean F_{ST} values of fine-mapped caQTL split by cell state. The dashed red line indicates the genome-wide average for non-caQTL SNVs. (**K**) Heatmap illustrating the proportion of caQTL-ACRs from each cell state that are temperate or tropical specific. Cell states with significant deviations from expected proportions (FDR < 0.05) are denoted by colored bubbles, where the size of the bubble indicates the odds ratio.

SHAPED COTYLEDON 2 (*ZmCUC2*), and *knotted 1* (*kn1*) were associated with shoot apical meristem (SAM) morphology (*40*), leaf width (*41*), and ear, tassel, and leaf architecture (*42*), respectively. CAWA deconvolution identified the TFBSs with the largest impact on organismal traits; SPL and homeodomain motifs associated with SAM and ear morphology, WRKY motifs with flowering time traits, and AUXIN RESPONSE FACTOR (ARF) motifs with leaf width (**fig. S14A**). Considering complex traits are manifest by cellular phenotypes across tissues and organs, we compared the percent of trait variation explained by cell-state-resolved fine-mapped caQTL relative to permutations of non-caQTL SNVs. Fine-mapped caQTL explained a significant proportion of trait variance for 96% (47/49) of complex traits (**Fig. 4B**). Additionally, organismal traits were influenced by multiple cell states, indicating cell-distinct processes impact phenotypic variation (**Fig. 4B**). However, fine-mapped caQTL from meristem and primordia cells explained the most variance for 69% (34/49) of traits (**fig. S14B**-**S14C**), including flowering time and SAM, ear, and tassel morphologies, supporting prior observations that adult architectural traits associate with regulators of SAM identity (*43*).

To better understand the genetic basis of trait variation, we integrated caQTL, eQTL, and GWAS hits through linkage analysis (r^2>0.5) of lead variants, revealing 4,395 variants connecting 444 caQTL-ACRs (1,285 unique caQTL SNVs), 295 eQTL-genes and 2,850 GWAS hits across 20 cell states (**table S11**). For example, an upper leaf angle GWAS hit was linked with an abaxial bundle sheath and phloem sieve element/procambial precursor caQTL and an eQTL for a cell wall-associated *ARABINOGALACTAN PROTEIN* gene (**Fig 4C**). Curiously, the GWAS variant was in stronger linkage with the caQTL than the eQTL. Comparing LD between molecular and phenotype-associated variants, we find caQTL are in stronger linkage with GWAS variants compared to eQTL globally (WRST; $P<2.2e^{-12}$; **fig. S14D**). Investigation of caQTL-GWAS trait linkage uncovered an enrichment of caQTL associated with flowering time traits (**fig. S14E**), leading us to hypothesize that most flowering

variability in the diversity panel is due to regulatory adaptation. Many physiological processes, including flowering, changed during the migration of maize from tropical to temperate latitudes. Comparing tropical and temperate maize lines, we found 10,740 ACRs with differential accessibility (FDR<0.1; hereafter referred to as subpopulation-specific ACRs; **Fig. 4D**). These subpopulation-specific ACRs were strongly enriched in signatures of selection that accompanied the adaptation to temperate climates and GWAS variants associated with flowering time (**Fig. 4E**; **fig. S14F**; **table S12**) (*44*). Moreover, different TFBSs were enriched in subpopulation-specific ACRs, including transcription-silencing motifs recognized by BPC TFs in tropical ACRs (*30*), and BRASSINAZOLE-RESISTANT 1 (BZR1), a master developmental regulator, AT-HOOK MOTIF CONTAINING NUCLEAR LOCALIZED 20 (AHL20), associated with delayed flowering (*45*), and a BASIC HELIX-LOOP-HELIX 18 (bHLH18) TF specifically expressed in floral tissue (*46*) in temperate ACRs (**Fig. 4F**; **table S13**). These TFBS enrichments are consistent with a modified floral transition being integral for temperate adaptation. Interestingly, despite including genetic relatedness covariates in caQTL mapping, subpopulation-specific ACRs were strongly coincident with caQTL-ACRs (**Fig. 4G**), suggesting that temperate/tropical adaptation underlies a significant proportion of chromatin accessibility variation. Indeed, allele frequency comparisons between tropical and temperate inbreds revealed higher F_{ST} values for caQTL and fine-mapped caQTL compared to non-caQTL SNVs (WRST, $P<2.2e^{-16}$; **Fig. 4H**).

To define the regulatory grammar promoting temperate and tropical differentiation, we compared TF footprints localized with fine-mapped caQTL categorized as high or low F_{ST}. High F_{ST} fine-mapped caQTL were enriched for TF footprints previously associated with flowering time (MYB56, IDD6) (*47*, *48*) and floral cell identity (DIV1) (*49*), and depleted of TF footprints associated with short vegetative phase (SVP) (*50*), ABA-responsiveness (ATHB-5) (*51*) and embryogenesis (NAC018) (*52*) (**Fig. 4I**; **table S14**). We hypothesized that local adaptation would impact chromatin accessibility in specific cell states. Indeed, F_{ST} distributions for fine-mapped caQTL differed among cell states and were heavily biased towards ACRs with increased temperate chromatin accessibility (Kruskal-Wallis Rank Sum Test, KWRST, $P<9.4e^{-12}$; **Fig. 4J-4K**). These results suggest that temperate adaptation involved widespread

changes to chromatin accessibility that rewired flowering and developmental TF networks and acted non-uniformly across cell states.

Discussion

Variable chromatin accessibility underlies cell-type-specific transcriptional regulation that drives functionally diverse cellular states. Although we previously established widespread chromatin accessibility variation among cell states within maize (*53*), the extent to which cell-type level chromatin accessibility differed between genetically diverse individuals remained limited. Here, we present a large-scale investigation of chromatin accessibility variation in 20 cell states across 172 geographically dispersed maize inbreds. Due to low LD in maize compared to humans or mice, we fine mapped >22,000 putatively causal variants. These fine-mapped caQTL are highly cell-state-specific and enriched in TCP TFBSs. The patterns of TCP gene expression, association with global co-accessibility, and significant effects on local chromatin accessibility collectively implicates TCP TFs regulators of chromatin accessibility. Identification of TCP TFs as prevalent and potent promoters of chromatin accessibility and contributors towards chromatin interactions adds to a growing body of evidence implicating TCP TFs as major regulators of chromatin architecture (*39*). This is a potentially exciting advance, given that only a handful of plant pioneer factors have been identified and the lack of a CTCF-binding factor homolog (a metazoan-specific chromatin looping factor) in all studied plants. We anticipate follow-up genetic and biochemical experiments to investigate the mechanistic roles of TCP factors in chromatin accessibility and long-distance chromosome looping.

Integrating caQTL with GWAS hits uncovered the cellular contexts associated with phenotypic traits. Further, the combination of CAWA, TWA and GWAS provides some preliminary insights into the underlying regulatory mechanisms behind complex trait variation. Examining the ACR differences between temperate and tropical maize, we demonstrated that caQTL are associated with signatures of population differentiation and selection, uncovering regulatory variation as a key target of local adaptation and human breeding efforts. As such, this resource provides exciting avenues for follow-up

exploration and direct real-world application. First, by identifying over 22,000 *cis* variants predicted to causally affect chromatin accessibility, ~1,300 of which are associated with quantitative phenotypes, we uncovered ideal molecular breeding and genetic engineering targets for crop improvement. Second, our work revealed how individual TFBS variants affect both the cell-type specificity and enhancer activity of TFBSs, which can fuel synthetic biological design. Third, we defined the ACRs and cognate CREs that are fixed in domesticated maize. The absence of extant genetic variation over these regions both highlights their agronomic importance and occludes their study through quantitative genetic techniques. As such, these fixed regions represent potent exploratory targets for engineering allelic variation via gene editing, both to understand their function and to shepherd novel crop phenotypes. Future reverse genetic and biochemical exploitation of these data will undoubtedly inform crop improvement and deepen mechanistic understanding of phenotypic variation.

Our study addressed several outstanding questions, serving as a foundation for similar caQTL studies. However, it profiles caQTL in only 20 of 33 cell states due to limited power in rare cell types. With technological increases in throughput, future studies will resolve the caQTL from low-frequency cell types. Additionally, we only sampled 7-day old seedlings; this tissue was selected because it includes the shoot apical meristem which is known to forecast adult traits. However, our sampling strategy cannot find caQTL specific to other phenotypically relevant organs including roots, inflorescences, and seeds. Finally, although the size of the diversity panel was powered to identify *cis*-QTL, it is limited for *trans*-QTL detection; uncovering these *trans*-regulatory interactions remains a challenge in most eukaryotic genomes. Future studies with larger cohorts and datasets that pre-identify *trans*-regulated ACRs provide a promising avenue to reveal the full complement of *trans*-QTL.

References and Notes

1. J. G. Wallace *et al.*, Association mapping across numerous traits reveals patterns of functional variation in maize. *PLoS Genet* **10**, e1004845 (2014).

2. J. Engelhorn *et al.*, Genetic variation at transcription factor binding sites largely explains phenotypic heritability in maize. *bioRxiv*, 2023.2008.2008.551183 (2024).

3. A. P. Marand, A. L. Eveland, K. Kaufmann, N. M. Springer, cis-Regulatory Elements in Plant Development, Adaptation, and Evolution. *Annual Review of Plant Biology* **74**, 111-137 (2023).

4. R. J. Schmitz, E. Grotewold, M. Stam, Cis-regulatory sequences in plants: Their importance, discovery, and future challenges. *Plant Cell* **34**, 718-741 (2022).

5. F. He *et al.*, Cis-regulatory evolution spotlights species differences in the adaptive potential of gene expression plasticity. *Nature Communications* **12**, (2021).

6. E. Z. Kvon *et al.*, Progressive Loss of Function in a Limb Enhancer during Snake Evolution. *Cell* **167**, 633-+ (2016).

7. C. Huang *et al.*, ZmCCT9 enhances maize adaptation to higher latitudes. *Proc Natl Acad Sci U S A* **115**, E334-E341 (2018).

8. S. Salvi *et al.*, Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. *Proc Natl Acad Sci U S A* **104**, 11376-11381 (2007).

9. A. Studer, Q. Zhao, J. Ross-Ibarra, J. Doebley, Identification of a functional transposon insertion in the maize domestication gene tb1. *Nat Genet* **43**, 1160-1163 (2011).

10. J. Tian *et al.*, Teosinte ligule allele narrows plant architecture and enhances high-density maize yields. *Science* **365**, 658-664 (2019).

11. C. J. Whipple *et al.*, grassy tillers1 promotes apical dominance in maize and responds to shade signals in the grasses. *Proc Natl Acad Sci U S A* **108**, E506-512 (2011).

12. Q. Yang *et al.*, CACTA-like transposable element in ZmCCT attenuated photoperiod sensitivity and accelerated the postdomestication spread of maize. *Proc Natl Acad Sci U S A* **110**, 16969-16974 (2013).

13. M. A. A. Minow, A. P. Marand, R. J. Schmitz, Leveraging Single-Cell Populations to Uncover the Genetic Basis of Complex Traits. *Annu Rev Genet* **57**, 297-319 (2023).

14. C. International HapMap, A haplotype map of the human genome. *Nature* **437**, 1299-1320 (2005).

15. C. C. Laurie *et al.*, Linkage disequilibrium in wild mice. *Plos Genetics* **3**, 1487-1495 (2007).

16. Y. J. Xiao, H. J. Liu, L. J. Wu, M. Warburton, J. B. Yan, Genome-wide Association Studies in Maize: Praise and Stargaze. *Molecular Plant* **10**, 359-374 (2017).

17. J. M. Noshay *et al.*, Stability of DNA methylation and chromatin accessibility in structurally diverse maize genomes. *G3 (Bethesda)* **11**, (2021).

18. W. A. Ricci *et al.*, Widespread long-range cis-regulatory elements in the maize genome. *Nat Plants* **5**, 1237-1249 (2019).

19. A. Husbands, E. M. Bell, B. Shuai, H. M. Smith, P. S. Springer, LATERAL ORGAN BOUNDARIES defines a new family of DNA-binding transcription factors and can interact with specific bHLH proteins. *Nucleic Acids Res* **35**, 6663-6671 (2007).

20. X. Chen *et al.*, SQUAMOSA promoter-binding protein-like transcription factors: star players for plant growth and development. *J Integr Plant Biol* **52**, 946-951 (2010).

21. C. Schwechheimer, P. M. Schroder, C. E. Blaby-Haas, Plant GATA Factors: Their Biology, Phylogeny, and Phylogenomics. *Annu Rev Plant Biol* **73**, 123-148 (2022).

22. P. K. Agarwal, P. Agarwal, M. K. Reddy, S. K. Sopory, Role of DREB transcription factors in abiotic and biotic stress tolerance in plants. *Plant Cell Rep* **25**, 1263-1274 (2006).

23. M. A. Khoso *et al.*, WRKY transcription factors (TFs): Molecular switches to regulate drought, temperature, and salinity stresses in plants. *Front Plant Sci* **13**, 1039329 (2022).

24. J. M. Noshay *et al.*, Assessing the regulatory potential of transposable elements using chromatin accessibility profiles of maize transposons. *Genetics* **217**, 1-13 (2021).

25. H. Zhao *et al.*, Proliferation of Regulatory DNA Elements Derived from Transposable Elements in the Maize Genome. *Plant Physiol* **176**, 2789-2803 (2018).

26. S. Yanagisawa, The Dof family of plant transcription factors. *Trends Plant Sci* **7**, 555-560 (2002).

27. F. Licausi, M. Ohme-Takagi, P. Perata, APETALA2/Ethylene Responsive Factor (AP2/ERF) transcription factors: mediators of stress responses and developmental programs. *New Phytol* **199**, 639-649 (2013).

28. C. Smaczniak, R. G. Immink, G. C. Angenent, K. Kaufmann, Developmental and evolutionary diversity of plant MADS-domain factors: insights from recent studies. *Development* **139**, 3081-3098 (2012).

29. K. A. G. Kremling *et al.*, Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. *Nature* **555**, 520-523 (2018).

30. J. Xiao *et al.*, Cis and trans determinants of epigenetic silencing by Polycomb repressive complex 2 in Arabidopsis. *Nat Genet* **49**, 1546-1552 (2017).

31. Y. Zhou *et al.*, Telobox motifs recruit CLF/SWN-PRC2 for H3K27me3 deposition via TRB factors in Arabidopsis. *Nat Genet* **50**, 638-644 (2018).

32. B. Wang *et al.*, Genome-wide selection and genetic improvement during modern maize breeding. *Nat Genet* **52**, 565-571 (2020).

33. P. Sijacic, M. Bajic, E. C. McKinney, R. B. Meagher, R. B. Deal, Changes in chromatin accessibility between Arabidopsis stem cells and mesophyll cells illuminate cell type-specific transcription factor networks. *Plant J* **94**, 215-231 (2018).

34. S. Miyashima *et al.*, Mobile PEAR transcription factors integrate positional cues to prime cambial growth. *Nature* **565**, 490-494 (2019).

35. G. Wu *et al.*, KANADI1 regulates adaxial-abaxial polarity in Arabidopsis by directly repressing the transcription of ASYMMETRIC LEAVES2. *Proc Natl Acad Sci U S A* **105**, 16392-16397 (2008).

36. S. M. de Jager *et al.*, Dissecting regulatory pathways of G1/S control in Arabidopsis: common and distinct targets of CYCD3;1, E2Fa and E2Fc. *Plant Mol Biol* **71**, 345-365 (2009).

37. L. Schlegel *et al.*, GenomicLinks: Deep learning predictions of 3D chromatin loops in the maize genome. *bioRxiv*, 2024.2005.2006.592633 (2024).

38. L. Dong *et al.*, Transcriptomic Analysis of Leaf Sheath Maturation in Maize. *International Journal of Molecular Sciences* **20**, 2472 (2019).

39. C. Liu, Y. J. Cheng, J. W. Wang, D. Weigel, Prominent topologically associated domains differentiate global chromatin packing in rice from Arabidopsis. *Nat Plants* **3**, 742-748 (2017).

40. J. W. Satterlee, J. Strable, M. J. Scanlon, Plant stem-cell organization and differentiation at single-cell resolution. *P Natl Acad Sci USA* **117**, 33689-33699 (2020).

41. R. Johnston *et al.*, Transcriptomic Analyses Indicate That Maize Ligule Development Recapitulates Gene Expression Patterns That Occur during Lateral Organ Initiation. *Plant Cell* **26**, 4718-4732 (2014).

42. R. A. Kerstetter, D. Laudencia-Chingcuanco, L. G. Smith, S. Hake, Loss-of-function mutations in the maize homeobox gene, knotted1, are defective in shoot meristem maintenance. *Development* **124**, 3045-3054 (1997).

43. S. Knauer *et al.*, A high-resolution gene expression atlas links dedicated meristem genes to key architectural traits. *Genome Res* **29**, 1962-1973 (2019).

44. H. Liu *et al.*, Genomic, Transcriptomic, and Phenomic Variation Reveals the Complex Adaptation of Modern Maize Breeding. *Mol Plant* **8**, 871-884 (2015).

45. R. Tayengwa, P. Sharma Koirala, C. F. Pierce, B. E. Werner, M. M. Neff, Overexpression of AtAHL20 causes delayed flowering in Arabidopsis via repression of FT expression. *BMC Plant Biol* **20**, 559 (2020).

46. A. V. Klepikova, A. S. Kasianov, E. S. Gerasimov, M. D. Logacheva, A. A. Penin, A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. *Plant J* **88**, 1058-1070 (2016).

47. L. Y. Chen, A. Bernhardt, J. Lee, H. Hellmann, Identification of MYB56 as a Novel Substrate for CRL3 E3 Ligases. *Molecular Plant* **8**, 242-250 (2015).

48. J. Colasanti, Z. Yuan, V. Sundaresan, The indeterminate gene encodes a zinc finger protein and regulates a leaf-generated signal required for the transition to flowering in maize. *Cell* **93**, 593-603 (1998).

49. D. Zhang *et al.*, Transcription factor DIVARICATA1 positively modulates seed germination in response to salinity stress. *Plant Physiol*, (2024).

50. U. Hartmann *et al.*, Molecular cloning of SVP: a negative regulator of the floral transition in Arabidopsis. *Plant J* **21**, 351-360 (2000).

51. H. Johannesson, Y. Wang, J. Hanson, P. Engström, The homeobox gene is a potential regulator of abscisic acid responsiveness in developing seedlings. *Plant Molecular Biology* **51**, 719-729 (2003).

52. T. Kunieda *et al.*, NAC Family Proteins NARS1/NAC2 and NARS2/NAM in the Outer Integument Regulate Embryogenesis in. *Plant Cell* **20**, 2631-2642 (2008).

53. A. P. Marand, Z. Chen, A. Gallavotti, R. J. Schmitz, A cis-regulatory atlas in maize at single-cell resolution. *Cell* **184**, 3041-3055 e3021 (2021).

54. X. Zhang, A. P. Marand, H. Yan, R. J. Schmitz, scifi-ATAC-seq: massive-scale single-cell chromatin accessibility sequencing using combinatorial fluidic indexing. *Genome Biology* **25**, 90 (2024).

55. M. B. Hufford *et al.*, De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science* **373**, 655-662 (2021).

56. H. Zhang *et al.*, Fast alignment and preprocessing of chromatin profiles with Chromap. *Nat Commun* **12**, 6566 (2021).

57. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

58. Picard2019toolkit, Picard toolkit. *Broad Institute, GitHub repository*, (2019).

59. R. Bukowski *et al.*, Construction of the third-generation Zea mays haplotype map. *Gigascience* **7**, 1-12 (2018).

60. E. Garrison, G. Marth, Haplotype-based variant detection from short-read sequencing. *arXiv e-prints*. 2012.

61. P. Danecek *et al.*, The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).

62. H. Heaton *et al.*, souporcell: Robust clustering of single cell RNAseq by genotype and ambient RNA inference without reference genotypes. *bioRxiv*, 699637 (2019).

63. I. Korsunsky *et al.*, Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods* **16**, 1289-1296 (2019).

64. J. B. Kang *et al.*, Efficient and precise single-cell reference atlas mapping with Symphony. *Nature Communications* **12**, 5890 (2021).

65. K. Zhang *et al.*, A single-cell atlas of chromatin accessibility in the human genome. *Cell* **184**, 5985-6001.e5919 (2021).

66. B. Nelms, V. Walbot, Defining the developmental program leading to meiosis in maize. *Science* **364**, 52-56 (2019).

67. Y. Zhang *et al.*, Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**, R137 (2008).

68. N. Kryuchkova-Mostacci, M. Robinson-Rechavi, A benchmark of gene expression tissue-specificity metrics. *Brief Bioinform* **18**, 205-214 (2017).

69. J. D. Welch *et al.*, Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177**, 1873-1887 e1817 (2019).

70. J. Armstrong *et al.*, Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**, 246-+ (2020).

71. S. C. Stelpflug *et al.*, An Expanded Maize Gene Expression Atlas based on RNA Sequencing and its Use to Explore Root Development. *Plant Genome* **9**, (2016).

72. K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, A. Siepel, Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**, 110-121 (2010).

73. G. Hickey, B. Paten, D. Earl, D. Zerbino, D. Haussler, HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* **29**, 1341-1342 (2013).

74. X. Zheng *et al.*, A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326-3328 (2012).

75. A. Taylor-Weiner *et al.*, Scaling computational genomics to millions of individuals with GPUs. *Genome Biology* **20**, 228 (2019).

76. O. Troyanskaya *et al.*, Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520-525 (2001).

77. C. C. Chang *et al.*, Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, (2015).

78. Y. Zou, P. Carbonetto, G. Wang, M. Stephens, Fine-mapping from summary data with the "Sum of Single Effects" model. *PLoS Genet* **18**, e1010299 (2022).

79. M. Bentsen *et al.*, ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nature Communications* **11**, 4267 (2020).

80. S. G. Coetzee, G. A. Coetzee, D. J. Hazelett, motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics* **31**, 3847-3849 (2015).

81. A. P. Marand, Computational Analysis of Maize Enhancer Regulatory Elements Using ATAC-STARR-seq. *Bio-101* **14**, e4953 (2024).

82. J. Ou, S. A. Wolfe, M. H. Brodsky, L. J. Zhu, motifStack for the analysis of transcription factor binding site evolution. *Nat Methods* **15**, 8-9 (2018).

83. I. Rauluseviciute *et al.*, JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **52**, D174-D182 (2024).

84. X. Liu, M. Huang, B. Fan, E. S. Buckler, Z. Zhang, Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. *PLoS Genet* **12**, e1005767 (2016).

85. K. Zhao *et al.*, Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. *Nat Commun* **2**, 467 (2011).

Acknowledgements

We wish to thank the GACRC for providing valuable computational support and resources, USDA GRIN for maize germplasm, the Duke University School of Medicine Sequencing and Genomic Technologies Shared Resource for Illumina sequencing services, Dr. Ed Buckler for advice on genotyping strategies, Dr. Joseph Gage for modeling approaches, and Dr. Nathan Springer for suggested analyses.

Funding:

National Institutes of Health grant 1R00/K99GM144742 (A.P.M.)

National Science Foundation grant IOS-1905869 (A.P.M.)

National Science Foundation grant IOS-2026554 (R.J.S.)

National Science Foundation grant MCB-2120132 (R.J.S.)

National Science Foundation grant IOS-1856627 (R.J.S.)

University of Georgia Office of Research (R.J.S.)

Author contributions:

Conceptualization: A.P.M., R.J.S.

Methodology: A.P.M., X.Z.

Investigation: A.P.M., L.J., F.C.G., M.A.A.M., J.P.M., Z.L., S.B., C.M., L.S., F.J., R.J.S.

Visualization: A.P.M., L.J., F.C.G.

Funding acquisition: A.P.M., R.J.S.

Project administration: A.P.M., R.J.S.

Supervision: A.P.M., R.J.S., F.J.

Writing – original draft: A.P.M., M.A.A.M., R.J.S.

Competing interests: R.J.S. is a co-founder of Request Genomics, LLC, a company that provides epigenomic services. The remaining authors declare no competing interests.

Data and materials availability: All data supporting the conclusions of this study have been made available within the article and the supplementary information. Raw and processed data generated by this study are available in the National Center for Biotechnology Information Gene Omnibus (NCBI GEO) under accession number GSE275410 and on Dryad (https://doi.org/10.5061/dryad.gxd2547vt). Coverage tracks from the different inbred lines can be viewed on JBrowse (https://epigenome.genetics.uga.edu/PlantEpigenome/?data=maize_v5). Scripts and software used in the analysis of the data generated by this study are publicly available at https://github.com/plantformatics/maize_282_diversity_scATAC.

Supplementary Materials

Materials and Methods

Figures S1-S14

Tables S1-S14 (separate files)

Supplementary Materials for

Title: The genetic architecture of cell-type-specific *cis*-regulation

Alexandre P. Marand[1†*], Luguang Jiang[2], Fabio Cano-Gomez[2], Mark A.A. Minow[1], Xuan Zhang[1], John P. Mendieta[1], Ziliang Luo[1], Sohyun Bang[3], Haidong Yan[1], Cullan Meyer[1], Luca Schlegel[4], Frank Johannes[4], Robert J. Schmitz[1*]

Correspondence to Alexandre P. Marand (amarand@umich.edu) and Robert J. Schmitz (schmitz@uga.edu)

This PDF files includes:

Materials and Methods
Methods references
Figures S1 to S14
Captions for Tables S1 to S14
References

Other Supplementary Materials for this manuscript include the following:

Data Tables S1 to S14

Materials and Methods

Plant material and growth conditions

Kernels for 290 maize genotypes were acquired from USDA ARS-GRIN (https://npgsweb.ars-grin.gov). Five biological replicates were sown in 1-inch plug trays with Sungro soil (Sungro Horticulture Canada Ltd) for each genotype in a walk-in growth chamber. Seedlings were maintained under 16-hour 50/50 mixture of 4100K (Sylvania Supersaver Cool White Delux F34CWX/SS, 34W) and 3000K (GE Ecolux w/ starcoat, F40CX30ECO, 40W) fluorescent lighting 50 cm above soil for 7 days at approximately 25°C during light hours with a relative humidity of approximately 54%. Of the 290 genotypes, 177 had germinated after 7 days, the remaining genotypes were excluded from the study. A single representative seedling was selected for each genotype and partitioned to 8 mm either side of the first visible node (16 mm total) with a #2 razor. The 16 mm leaf roll section were rinsed with deionized water and pooled according to the experimental design in **table S2**. All samples were collected between 8 and 9 AM.

Single-cell ATAC-seq library construction

For each multiplexed pool, leaf roll sections were suspended in 500 uL of LB01 buffer (15mM Tris pH 7.5, 2mM EDTA, 0.5mM Spermine, 80mM KCl, 20mM NaCl, 15mM 2-mercaptoethanol, 0.15% TrixtonX-100) and homogenized with a No. 2 straight-edge razor for 120 seconds. The lysate was filtered through two layers of miracloth (Millipore #475855), stained with DAPI to a final concentration of ~ 1uM and loaded onto a Beckman Coulter Moflo XDP flow cytometer instrument. Between 60,000-120,000 nuclei were sorted from each pool into 1.5mL Eppendorf LoBind® centrifuge tube containing ~175-350 uL of LB01. Sorted nuclei were spun in a swinging bucket centrifuge (5 minutes, 500 rcf) and resuspended in 10 uL of LB01, visualized and counted on a hemocytometer with a fluorescence microscope, and adjusted to a final concentration between 3,200-12,000 nuclei per uL using diluted nuclei buffer (DNP; 10X Genomics, catalog: 1000176). For each pool, 5 uL of suspended nuclei were loaded per well on the Next GEM Chip H (10X Genomics, catalog: 1000162). Single-cell ATAC-seq libraries were prepared according to the manufacturer's instructions (10X Genomics, catalog: 1000176,

Chromium Next GEM v1.1) using the Chromium Controller (10X Genomics, catalog: 120223). scATAC-seq libraries were sequenced using an Illumina S4 flow cell (NovaSeq 6000) in dual-index mode with eight and 16 cycles for i7 and i5, respectively.

Single-nucleus RNA-seq library construction

The protocol for nuclei isolation and purification was adapted from the previously described scifi-ATAC-seq method (*54*). In summary, to minimize RNA degradation and leakage, the same tissue used for scATAC-seq from two maize genotypes, B73 and Mo17, was finely chopped on ice for approximately 1 minute using 600 μL of pre-chilled Nuclei Isolation Buffer containing 0.4U/μL RNase inhibitor (Roche, Protector RNase Inhibitor, Cat. RNAINH-RO) and a comparatively low detergent concentration of 0.1% NP-40. Following chopping, the entire mixture was passed through a 40-μm cell strainer and then subjected to centrifugation at 500 rcf for 10 minutes at 4°C. The supernatant was carefully decanted, and the pellet was resuspended in 500 μL of NIB wash buffer, comprising 10 mM MES-KOH at pH 5.4, 10 mM NaCl, 250 mM sucrose, 0.5% BSA, and 0.2U/μL RNase inhibitor. The sample was filtered again, this time through a 20-μm cell strainer, and gently layered onto the surface of 1 mL of a 35% Percoll buffer. The Percoll buffer was prepared by mixing 35% Percoll with 65% NIB wash buffer in a 1.5-mL centrifuge tube. The nuclei were then subjected to centrifugation at 500 rcf for 10 minutes at 4°C. After centrifugation, the supernatant was carefully removed, and the pellets were washed twice in 650 μL of Nuclei Wash & Resuspension buffer (NWR buffer: 1% BSA, 0.2U/μL RNase inhibitor, 1xPBS). The nuclei were resuspended in 50 μL NWR buffer. Approximately 5 μL of nuclei were diluted tenfold and stained with DAPI (Sigma Cat. D9542). Subsequently, nuclei quality and density were evaluated with a hemocytometer under a fluorescence microscope. The original nuclei were further diluted with DNB buffer to achieve a final concentration of 1,000 nuclei/μL. Ultimately, a total of 16,000 nuclei were used as input for snRNA-seq library preparation.

For scRNA-seq library preparation, we employed the Chromium Next GEM Single Cell 3'GEM Kit v3.1 from 10X Genomics (Cat# PN-1000123), following the manufacturer's instructions (10xGenomics,

CG000315_ChromiumNextGEMSingleCell3-_GeneExpression_v3.1_DualIndex_RevB). The libraries were subsequently sequenced using the Illumina NovaSeq 6000 in dual-index mode with 13 cycles for the i7 and i5 indices, respectively.

scATAC-seq raw data processing

scATAC-seq binary base call sequences files (BCL) output from the Illumina S4 NovaSeq 6000 were demultiplexed with *cellranger-atac mkfastq* (v1.2) and aligned to the B73 AGP v5 maize reference genome (55) using *chromap* (v0.2.3) with non-default settings (-l 2000 -q 1 --low-mem --trim-adapters --barcode-whitelist --bc-error-thresholds 2 --remove-pcr-duplicates-at-cell-level) (56). Only uniquely and properly paired reads with mapping quality greater than 10 were retained using *samtools view* (v1.16.1) with non-default settings (-bhq 10 -f 3) (57). Duplicated fragments were removed on a per-barcode basis using *picardtools MarkDuplicates* (v2.27.4) with non-default settings (BARCODE_TAG=CB REMOVE_DUPLICATES=true) (58). Alignments were then used to extract single-base resolution Tn5 integration sites by shifting the start and end coordinates by +5 and -4, respectively. Tn5 integration sites were filtered to retain only unique coordinates for each barcode.

Variant processing

To partition reads and barcodes by their respective genotypes, we first re-aligned short read whole genome sequencing (WGS) data from all 172 inbred lines (59) to the maize B73 v5 reference genome (55) using *BWA mem* (v0.7.17) with non-default settings (-M -t 4). Inbred lines with more than one WGS experiment were merged into a single BAM file. WGS alignments for each inbred line were then sorted and deduplicated using *samtools* (v1.16.1) (57). Variants were identified using *freebayes* (v1.2.0) with non-default settings (--strict-vcf -n 4 -q 10 --max-coverage 10000) (60). To simplify downstream analysis, we only retained bi-allelic SNPs. Genotype calls with fewer than four reads were set to missing, and variants with more than 25% missing data or a minor allele frequency less than 5% were removed, resulting in a final set of 31,478,930 bi-allelic SNPs. Weir and Cockerham F_{st} was estimated

using *vcftools* (v0.1.16) with non-default parameters (--weir-fst-pop $tropical –weir-fst-pop $temperate) (*61*).

In silico genotype-demultiplexing

To assign barcodes to their genotype(s) of origin, we first removed barcodes with fewer than 100 unique Tn5 integration sites. Genotype and singlet/doublet probabilities for individual barcodes were estimated using *souporcell* with non-default values (-k $num_geno --min-alt 5 --min-ref 5 --skip-remap TRUE --no-umi TRUE) where $num_geno references the number of genotypes multiplexed in the sequencing library (*62*). Barcodes with a greater singlet to doublet probability and the best genotype assignment with 1000-fold greater likelihood than the next best genotype assignment were retained for downstream analysis.

scATAC-seq quality control

Starting from the unfiltered data set, barcodes with at least 500 unique Tn5 integration sites, standardized (Z-score) fraction of Tn5 integration sites within 2-kb of the TSS greater than the maximum of two standard deviations below the mean and 0.2, and standardized (Z-score) fraction of Tn5 integration sites within ACRs greater than the maximum of two standard deviation below the mean and 0.2, were identified as putative nuclei. To further remove barcodes representing broken nuclei or ambient background, we first used the unfiltered set of barcodes to create a binary sparse matrix of barcodes x 500-bp bins tiled across the maize genome. To determine if a barcode's genome-wide chromatin profile was indicative of an intact nucleus or background noise, we estimated Spearman's correlation coefficients for each barcode with the following aggregates normalized by term frequency inverse document frequency (TF-IDF): (*i*) top 1,000 barcodes ranked by Tn5 insertion counts (non-background) and (*ii*) barcodes with less than 100 Tn5 insertions (background). Barcodes with greater correlations to the non-background aggregate relative to the background aggregate were retained.

scATAC-seq clustering using reference nuclei

To reduce computational costs and improve cell state identification, we first isolated a reference set of barcodes by removing nuclei from the filtered barcode x 500-bp bin sparse binary matrix with less than 1,000 accessible features using the *cleanData* function of *Socrates* (53) with non-default parameters (min.c=1000, min.t=0.0025, max.t=0). As a result, bins with chromatin accessibility frequency of 0.25% or greater across barcodes were retained. The binarized sparse matrix was then transformed using term frequency inverse document frequency (TFIDF) and L2 normalization in *Socrates* using the function *tfidf* with non-default parameters (doL2=T). The dimensionality of TFIDF normalized matrix was reduced to 21 scaled Principal Components (PCs) using Singular Value Decomposition (SVD) using the top 25% of variable features, with the first PC removed due to high correlation (PCC > 0.7) with nuclear read depth via the function *reduceDims* with non-default parameters (method="SVD", n.pcs=21, cor.max=0.7, num.var=ceiling(nrow(obj$counts)*0.25), scaleVar=T, doL2=T). Library and genotype effects were removed using *Harmony* (63) with non-default parameters (theta=c(2,2), tau=5, sigma=0.1, lambda=c(0.1, 0.1), nclust=50, max.iter.cluster=100, max.iter.harmony=30), return_object=T). The integrated object was used to create a reference embedding using the R package, *symphony* (64), via the *buildReferenceFromHarmonyObj* function with non-default parameters (do_umap=F, umap_min_dist=0.01). To visualize chromatin accessibility similarity among reference nuclei, we further reduced the dimensionality of the PC matrix using *umap* with non-default parameters (metric="cosine", n_neighbors=30, a=2, b=0.75, ret_model=T). The resulting model was added to the symphony reference object while the UMAP embedding was added to the UMAP slot of the *Socrates* object. Clusters of nuclei with similar chromatin accessibility patterns were identified using the *callClusters* function of *Socrates* with non-default parameters (res=$optres, k.near=30, cleanCluster=F, cl.method=4, e.thresh=3, threshold=3, m.clst=50) using the Leiden graph-based clustering algorithm. The optimum resolution for Leiden clustering was determined by iterating over a range of resolution parameters (0 – $5e^{-3}$) and estimating clustering stability scores at each resolution. Cluster stability was estimated perturbing 2% of graph edges and taking the clustering consensus values across 5 permutations, as previously described (65). The resolution parameter with the greatest cluster stability score was selected for the final clustering call.

To project the remaining cells on to the reference embedding, we first filtered nuclei with fewer than 100 accessible features using the function *cleanData* with non-default parameters (min.c=100, min.t=0, max.t=0). Feature IDs from the query nuclei were then aligned with the reference object to enable downstream integration. The remaining query nuclei were projected into the same TFIDF space as the reference nuclei using the function *projectTFIDF* with default parameters. Nuclei were then split into groups of ~5,000 and iteratively projected onto the reference SVD embedding. Specifically, during each iteration of new nuclei, we ran *mapQuery2*, a modified scATAC-seq variant of the *mapQuery* from *symphony*, using non-default parameters (svd_diag=objPCA_modeld[objPCA_modelkeep_pcs], vars=c("library", "genotype", sigma=0.01, do_umap=T, scaleVar=F, doL2=T).

scATAC-seq cell-type annotation

To annotate cell types, we first estimate gene activity scores by counting unique Tn5 integration sites overlapping gene bodies and in proximal regions 500-bp upstream of TSSs using the *Socrates* function, *estGeneActivity*. Gene activity scores were scaled to sum to 10,000 per nucleus. Cell type annotation was then performed using *a priori* marker gene activity UMAP profiles, cell type marker enrichment, differential gene activity, and multinomial logistic classification as previously described (*53*). Additionally, due to the effects of cell cycle stage towards clustering and interpreting cell type annotations, we first collected a list of known cell cycle stage-specific genes (*66*). We then estimated the enrichment of genes specific to each stage (average normalized gene activity) versus 1,000 permutations of the same number of randomly selected cell cycle genes within each nucleus. Final cell cycle stage scores (Z-scores) were derived by standardizing the average normalized gene accessibility of the focal stage by subtracting and dividing the mean and standard deviation of the permuted enrichment scores, respectively. The cell cycle stage with the highest normalized score was used as the predicted cell cycle stage. Final cell state annotations were generated by considering collection of predicted cell type and cycle stage information.

ACR identification

A comprehensive set of ACRs were first identified using aggregated Tn5 insertion counts for each genotype x cell type combination using MACS2 (v2.2.7.1) (67) with non-default parameters (-g 1.6e9 --nomodel --keep-dup all --extsize 150 --shift -75 --qvalue 0.05). ACRs from all genotype by cell-type combinations were then concatenated into a single master list (n=65,450,651). To address the presence of overlapping intervals without merging, we recursively selected the most significant ACR (based on $-\log_{10}$ q-value) and removed all ACRs with lesser $-\log_{10}$ q-values and direct overlap with the focal ACR, repeating the process iteratively to produce a final set of non-overlapping ACRs (n=4,323,286). For each cell type, we retained ACRs from the non-overlapping set that were supported by ACR calls (intersection >= 1bp) from at least two distinct genotypes, resulting in 631,846 ACRs. To remove ACRs with chromatin accessibility enrichment similar to the background Tn5 insertion rate, we collected the same number of random genomic regions (n=631,846) matched by interval size specifically from mappable regions. Mappable regions were defined as genomic regions covered by at least one read from a simulated data set generated by *wgsim* (57) with non-default parameters (-N 5e8 -1 100 -2 100 -d 300). We then constructed binary sparse matrices for (1) control regions x cell and (2) ACRs x cell and estimated the average Tn5 insertion frequency for each feature across all cells in each matrix. The average Tn5 insertion frequency across cells for each control region was then used to fit a beta distribution with *fitdist* from the *fitdistrplus* R package. We obtained *P*-values for each ACR by comparing the observed Tn5 insertion frequency to the null model parameters derived from control regions using the base R function *pbeta*. *P*-values were corrected for multiple testing using *p.adjust* with non-default parameters (method="fdr"). ACRs with an FDR less than 0.01 were retained, resulting in a total of 108,843 ACRs. For analyses specific to *cis*-regulatory function, we estimated the proportion of each ACR corresponding to exonic, intronic, intergenic (>2-kb upstream of the TSS), 50-bp surrounding the TSS, and proximal (<2-kb upstream of the TSS) sequences. We then removed all ACRs where the majority of the ACR interval overlapped an exon, except in cases where an ACR also overlapped with a TSS. Altogether, we identified a total of 82,098 ACRs with putative *cis*-regulatory function for downstream analysis. ACRs were defined as conserved if at least 75% of genotypes

exhibited at least 90% shared sequence alignment with B73. Cell-state specificity was estimated via the Tau metric (68) and ACRs with Tau greater than 0.7 were considered as cell-state-specific.

snRNA-seq raw data processing

We re-analyzed one previously generated snRNA-seq dataset (NCBI GEO accession number: GSM5344025) from the same tissue type in conjunction with the three newly generated snRNA-seq data (53). For the new data sets, binary base call sequences files (BCL) output from the Illumina S4 NovaSeq 6000 were demultiplexed with *cellranger mkfastq* (v6.1.1). Now using all new and previously generated data sets, raw reads were mapped with *cellranger count* resulting in BAM files containing all reads (both mapped and unmapped) with whitelist-corrected barcodes in the CB tag and UMI sequences in the UM tag. Alignment files were then filtered to remove low mapping quality alignments with *samtools view* (v1.16.1) (*-q 30*) and to remove duplicate reads using the *picardtools* (v2.27.4) function, *MarkDuplicates* by setting the BARCODE_TAG parameter equal to the concatenation of the CB and UMI sequences residing in a custom tag labeled UM. We then converted the filtered BAM alignments into BED-formatted files by moving the cell index and UMI sequence to the read name and piping the output to *bedtools bamtobed* (v2.27.1). Read counts per gene per cell were estimated by counting the overlap of alignments with gene bodies (including introns) for each gene via a custom perl script. Barcodes with fewer than 750 UMIs and 250 transcribed genes, and greater than 10,000 transcribed genes were removed. Barcodes with more than 5% of UMIs mapping to chloroplast or mitochondrial genes were also discarded, retaining barcodes where 95% of UMI were derived from the nuclear genome. We also filtered cell barcodes with gene expression patterns more correlated to background transcript profiles, similar to scATAC-seq data processing (see above). Cell genotypes were identified using *souporcell* (62) with non-default parameters (--threads 30 -k 2 --known_genotypes $vcf --known_genotypes_sample_names Mo17 B73), with barcodes identified as doublets removed from the analysis.

Multimodal data integration and cell-type annotation

Gene chromatin accessibility and transcript abundance matrices were first filtered to retain genes accessible/transcribed in at least 0.1% of nuclei (and nuclei with at least 100 accessible/transcribed genes) and matching between modalities. The ATAC and RNA matrices were loaded into the R package, *rliger* (69), and normalized with the function *normalize* with default parameters. The top most variable genes were selected using the function *selectGenes* with non-default parameters (datasets.use="RNA", var.thresh=0.5) and the matrices from both modalities were scaled without centering using the function *scaleNotCenter* to preserve the non-negative distributions expected by non-negative matrix factorization. Joint non-negative matrix factorization was performed with the function *optimizeALS* with non-default parameters (k=20, lambda=5). Finally, we performed quantile normalization using *quantile_norm* with default parameters to fully integrate the cell factor loadings from the two modalities. Nuclei clustering was performed similarly to the scATAC-seq data with the exception of setting k=15 and using 20 cell factors loadings returned by quantile normalization when calling clusters with the consensus-based approach. Cell type annotation was performed identically as described for the scATAC-seq only data set, including visualization of marker genes via UMAP and cluster-aggregated heatmaps, estimation of differentially accessible and transcribed genes, and automated per-cell annotations via multinomial regression.

phyloP score generation

Generation of phyloP scores was done in a multi-step process. Genome alignments for 12 monocot genomes, including *Zea mays, Sorghum bicolor, Oryza sativa, Brachypodium distachyon, Urochloa fusca, Panicum halli, Hordeum vulgare, Panicum miliaceum, Setaria itatlica, Eleusine coracana, Saccharum spontaneum, Miscanthus sinensis* was done using progressive *cactus* (version 2.6.7) (70) on softmasked genomes via RepeatMasker. After whole genome alignment, neutral sites were identified in the *Zea mays* genome to act as a null model. To do so, functional genomic data in the form of ChIP-seq peaks associated with functional markers correlated with transcription (H3K36me3, H3K4me1, H3K4me3, and H3K56ac from (18)) were removed. Additionally, regions of the genome with evidence of transcription (greater than 5 RNA-seq reads from reference (71) aligning to any given

locus), as well as regions overlapping exons and introns were removed. All neutral sites were required to additionally be found in at least 9 of the provided genomes.

The isolated neutral sites were then aggregated together and run through *phyloFit* (version 1.5) (*72*) with the default options including `--subst-mod REV --msa-format MAF`. The resulting model "mod" file as well as the location of the neutral sites were then used as input for the hal implementation of phylop, halTreePhyloP.py version 2.0 with defaults (*73*). The resulting bedgraph files were used for downstream analysis.

Preprocessing for caQTL mapping

Prior to running caQTL analysis, we first restricted the ACR set to regions where genotype and/or cell type was significant (FDR < 0.05) in a generalized logistic regression model that included (a) $\log_{10}$-transformed number of accessible sites, (b) sequencing library ID, (c) genotype, and (d) cell cycle stage. For identification of cell-type-associated ACRs, we also included the first five principal components from PCA of 31M bi-allelic SNPs. For identification of genotype-associated ACRs, we included the first two components from the symphony embedding. A total of 57,984 non-exonic ACRs varied significantly among genotypes, cell types, or both were identified as candidates for caQTL mapping. To estimate the relative variance explained by cell type and genotype, we ran *fitExtractVarPartModel* from the R package, *variancePartition* using cell type and genotype as random effects and all covariates as fixed effects.

To enable efficient caQTL mapping, we adopted a pseudobulking approach by aggregating Tn5 insertions of cells corresponding to the same cell type and genotype across all ACRs. The raw counts matrix was scaled by counts per million (CPM), quantile normalized, and subjected to inverse normalization per ACR to mitigate the effects of outliers. After partitioning the *genotype-cell type* (columns) x *ACR* (rows) matrix by cell type, we ran PCA for each distinct cell type matrix composed of *genotype* x *ACR* with CPM-scaled and ACR-standardized counts. Simultaneously, we estimated the top

10 PCs from the 31,478,930 bi-allelic SNVs used to genotype individual cells using the *R* package, *SNPRelate* (74). For each *cell type* x *genotype* combination, we also collected information on the total number of Tn5 insertions within ACRs, the total number of cells, the fraction of cells derived from all sequencing libraries, and the fraction of cells corresponding to G1, G1/S, S, G2, or M phases of the cell cycle. Technical covariates were subjected to PCA to identify orthogonal sources of variation, with the top X PCs explaining 95% of the variation (depending on cell type) merged with the first 20 cell-type pseudobulk chromatin accessibility PCs and first five genetic PCs.

caQTL mapping

For identification of *cis*-associated genetic variants in bulk tissue and each individual cell type, we used *tensorQTL* (75) with a max *cis* distance of 1 Mb, which has been previously shown to be a conserved threshold given the rapid link disequilibrium decay ($r^2 < 0.03$) in this diversity panel. Prior to caQTL mapping, we stringently filtered the bi-allelic SNP data set to remove sites with MAF less than 0.05, heterozygosity greater than 5%, and missingness greater than 25%. Genotype likelihoods were used to determine allele dosages. All missing values were then imputed using the nearest neighbor via the *R* package, *impute* (76), via the function *impute.knn* with non-default parameters (k=10) for each chromosome independently. Finally, the imputed SNV set was converted to *plink2* (77) format with non-default parameters (--make-pgen --output-chr chrM --vcf $VCF dosage=DS). To identify caQTL, we ran *tensorQTL* with non-default parameters (--mode cis --covariates $cov --qvalue_lamba 0.85 --dosage). To retrieve all nominal p-values for downstream analysis, we ran *tensorQTL* with non-default parameters (--covariates $cov --dosages --mode cis_independent). We set a dynamic FDR (between 0.01 and 0.1) across cell types based on a linear model fit to the average Tn5 integration depth per genotype, where more stringent FDR thresholds are set for cell types with more sequencing information.

Identification of eQTL was performed similarly as caQTL with the exception of independent construction of covariates, which was conducted identically as the scATAC-seq data. Briefly, we aggregated the top

20 PCs from the standardized genotype x gene matrix with the top five genetic PCs based on the SNV genotypes from *SNPRelate*. The $\log_2$ CPM normalized genotype x gene transcript abundance matrix was obtained from a prior study (29). Gene IDs were converted to v5 names using the liftover function of *CrossMap*. Gene expression was inverse quantile normalized per gene to mitigate the effects of outliers prior to eQTL mapping. We used identical functions and parameters of *tensorQTL* to identify eQTL as described for caQTL above.

Fine-mapping caQTL

Statistical fine-mapping of caQTL was performed using *susieR* (78). Briefly, we first identified all SNVs within caQTL-ACR coordinates and extracted the caQTL summary statistics output from *tensorQTL*. We then ran *susie_rss* with non-default parameters including the SNV correlation matrix adjusted by the same covariates given to *tensorQTL,* the beta hat and beta hat standard error from *tensorQTL* summary statistics*,* and set *L*=10 and *estimate_residual_variance*=T. Finally, we kept all SNVs within the 95% credible set list (the list of variants with 95% probability that includes the causal variant) with posterior inclusion probabilities greater than zero and FDR values less than the dynamic threshold described above*.*

Differential allelic TF footprinting

To identify TF footprints differing between alleles, we identified all SNVs within and flanking ACRs up to 1-kb away. We then identified the top two most frequent haplotypes of each ACR by clustering genotypes by their SNV dosage profiles using the R package, *mclust* with non-default parameters (G=1:10). Genotype membership for each haplotype was recorded across all ACRs, retaining ACRs with at least 5 individuals represented in the top two most frequent haplotype groups. Then, we aggregated scATAC-seq alignments for each ACR and each cell type to create pseudobulk aggregates reflecting the top two distinct haplotypes at cellular resolution. These alignments were then used as input for *TOBIAS* (79) along with custom reference FASTAs generated by *FastaAlternateReferenceMaker* from the *GATK* suite of tools for each haplotype. *TOBIAS* was run with

default parameters for the subfunctions, *ATACorrect* and *ScoreBigwig. TOBIAS BINDetect* was run with non-default parameters to increase the stringency in motif identification and footprint calling (--motif-pvalue 1e-5 --bound-pvalue 1e-4). We set an additional filter where differential TF footprints were then identified as regions with footprint scores greater than 5 in at least one haplotype and "unbound" based on *TOBIAS* thresholds in the other haplotype (footprint score bound/unbound threshold range across cell types = 0.95 - 1.58).

TF motif variant effects

Comparison of alleles overlapping predicted TFBSs while considering TF motif position weight matrices has been previously used to predict allelic effects on TF binding affinity (*80*). To disentangle the potential mechanisms of chromatin accessibility variation, we ran *motifbreakR,* an R package designed to estimate allelic effects on TF binding affinity (*80*). Specifically, we loaded the same variants used for caQTL mapping reformatted in BED as required by *motifbreakR,* as well as the coordinates of all caQTL-ACRs from cell state-resolved caQTL mapping. The SNV list was filtered to retain records embedded within the test set of ACR coordinates. We then ran *motifbreakR* using all plant-derived JASPAR 2022 motifs from the *MotifDb* R package with non-default parameters (filter=T, threshold=1e-5, method="ic", show.neutral=T).

Fine-mapping and TF footprint STARR-seq validation

To estimate enhancer activity of each allele for TF footprints containing fine-mapped variants, we first collected the reference and alternate sequences of all TF footprints containing fine-mapped variants. We then identified all reference and alternate allele TF footprint sequence occurrences in the genome for regions containing at least one mapped DNA or RNA STARR-seq read using *seqkit locate* (v2.5.1). Enhancer activity was estimated as previous described (*81*), which was used to map (via *bedtools map*) the maximum enhancer activity onto all reference and alternate TF footprint sequence occurrences. The enhancer activity of a given fine-mapped TF footprint allele was taken as the average enhancer activity across all occurrences genome wide, requiring at least one occurrence for both reference and

alternate alleles for a given fine-mapped TF footprint. We then estimated the allelic effect as the difference in averaged enhancer activity between alleles. For fine-mapped TF footprints with multiple genome wide occurrences for each allele (16,646 fine-mapped TF footprints), we further determined if alleles exhibited significant differences in enhancer activity using the Wilcoxon Rank Sum Test and correcting the nominal *P*-value using the Benjamini-Hochberg correction and kept records with FDR < 0.05.

Co-accessibility QTL mapping

To identify genetic variants putatively affecting higher-order chromatin interactions, we first estimated co-accessibility on a per-genotype basis. Briefly, we collected all cells (*n*) from each genotype in isolation, constructed pseudocells with the k-nearest neighbors (where *k* is the square root of *n*) and scaled Tn5 integration counts per ACR by counts per million. Raw co-accessibility scores (*A*) between all ACRs < 500-kb and > than 2-kb apart were then estimated using Spearman's correlation coefficient. We then penalized raw co-accessibility scores by the distance between ACRs to reflect the exponential decay observed empirically between pairs of interacting genomic regions using equation (eq) 1:

1. $A_{pen} = A \times \frac{\sqrt{2000}}{\sqrt{D}}$

Where *D* is the distance between the pair of ACRs. Finally, we estimated residuals between the penalized co-accessibility scores and a null set of permuted penalized co-accessibility scores derived from randomizing the cell x ACR matrix that maintained the same number of accessible cells per ACR. The set of co-accessible ACRs found across all genotypes was then filtered to retain links supported by previously published maize HiC and HiChIP chromatin loops (*18*). Genetic variants greater than 5Mb away from the focal co-accessible ACRs were then identified using *tensorQTL* with non-default parameters (--mode trans --covariates $cov --qvalue_lamba 0.85 --dosages --pval_threshold 0.0001 --batch_size 1000). Because transQTL mapping results in a substantially larger test space, we opted to focus on the number of traits associated with a single variant at $P<1e^{-5}$ rather than using multiple test

correction. Candidate genes were identified by first selecting all *trans*-QTL variants associated with more than 30,000 co-accessible ACRs and collecting gene IDs within 50-kb of the top variants.

Long-range caQTL analysis

To identify caQTL that potentially influence accessibility at other loci, we first identified caQTL-ACRs containing local caQTL (caQTL within the ACR coordinates). We then kept all caQTL records that were associated with two or more ACRs. The set of randomly permuted ACR-ACR interaction was derived by randomly linking ACRs from the long-range caQTL-ACR data set (i.e. scrambling the association between two ACRs using the same set of observed ACRs). Predicted interactions were determined by entering the long-range caQTL-ACR coordinates in BEDPE into the GenomicLinks web browser (*37*).

To assay the distribution of TF motifs among co-accessible ACRs with the support of caQTL evidence, we organize co-accessible ACRs into two main groups for analysis. The first group, our test set, includes co-accessible ACRs that have at least one caQTL and evidence from HiC/HiChIP experiments (*18*) showing they physically interact. The second group, our control set, consists of co-accessible ACRs showing physical interactions in HiC/HiChIP data but lacking caQTL support. To evaluate the enrichment of specific TF motifs within these groups, we employed the Fisher's Exact Test for Count Data in R (function: fisher.test, alternative="greater", R version 4.3.3). The contingency table was defined as follows: *matrix(c(n_i, d_i, c_i, t_i), nrow = 2)*. Here, 'n_i' represents the total number of co-accessible ACRs with a TF motif overlap, excluding those supported by both caQTL and HiC/HiChIP data; 'd_i' denotes co-accessible ACRs with HiC/HiChIP evidence but lacking caQTLs where the TF motif *i* is present; 'c_i' accounts for co-accessible ACRs supported by caQTL but not HiC/HiChIP; and 't_i' capture co-accessible ACRs supported by both caQTL and HiC/HiChIP data where the TF motif *i* is found. The detection of TF motifs in co-accessible ACRs was performed by scanning the corresponding PWM across all co-accessible ACRs. We only counted hits where the PWM was found in both anchors of the co-accessible ACR. The scanning of TF motifs was carried out with the *Motifmatchr* package in R (v1.24.0) with default settings. To minimize the redundancy of TF motifs arising from different

experiments of the same TF or from TFs within the same family with similar TF motifs, we aligned all PWMs and extracted their PWM signatures using the *clusterMotifs* and *motifSignature* functions (with parameters: cutoffPval < 0.0001 and min.freq = 1) from the *R* package *motifStack* (v1.46.0) (*82*). Finally, all PWM were collected from JASPAR 2024 keeping only PWM reported for *Arabidopsis thaliana* (*83*).

Complex trait association analyses

Agronomic traits of maize were downloaded from https://www.panzea.org/phenotypes. The genome-wide association analysis (GWAS) was conducted using the fixed and random model circulating probability unification (FarmCPU) approach, implemented in the R-based software *rMVP* (version 1.0.8; https://github.com/xiaolei-lab/rMVP) (*84*). The FarmCPU analysis was run using maxLoop = 10 and method.bin = "FaST-LMM". The first three genetic PCs were included in this model to correct the hidden population structure. To enable a comparison with the caQTL results, we exclusively retrieved the association of SNVs located within a 1Mb region upstream and downstream of ACRs (containing at least one caQTL) with various traits, resulting in an average of 37,449 SNP-trait associations (ranging from 608 to 67,970) being considered for each ACR. The significance threshold to control the type I error rate was corrected using the Bonferroni correction (adjusted $P < 1/n$) with respect to the number of SNVs across the ~2 Mb window. To estimate the phenotypic variance explained by each SNV, ANOVA was used with contrasting linear models eq. 2 and eq. 3:

2. $Y = \alpha X + \beta P + \varepsilon$
3. $Y = \beta P + \varepsilon$

Where Y is the phenotype, X is the SNV genotype, P is the PCs, α is the SNV effect, β is the PC effects, and ε is random error, which follows a $N(0, \sigma^2_e)$ distribution. Thus, the variance due to each SNV was reported after adjusting for the population structure effects, as previously described (*85*). We estimated the significance of the proportion of variance explained (PVE) of caQTL versus non-caQTL

by first partitioning caQTL by cell state and randomly selecting the same number of non-caQTL SNVs. We compared the mean SNV PVE for caQTL and non-caQTL SNVs across 10,000 permutations. For the transcriptome-wide association study (TWAS) and chromatin accessibility-wide association study (CAWAS), we re-distributed the normalized bulk CPM value for each gene and each ACR into a consecutively numeric range from 0 to 2, with the implementation of mixed linear model (MLM) in *rMVP*, using a relatedness matrix based on gene expression and ACR coverage intensity, respectively, to correct the population structure. Multiple testing corrections were performed with a false discovery rate (FDR) of 0.05 using the R function *p.adjust* (method = "fdr"). The relative contribution of individual TF motifs towards phenotypes was determined by matrix multiplication of the CAWAS results with a binarized motif x ACR matrix, scaling the subsequent motif scores from 0 to 1 across all motif for each trait.

Linkage analysis

The most significant SNV for the same ACR, expression and agronomic trait was regarded as lead SNV. The pairwise linkage disequilibrium (LD) between the lead caQTL, eQTL, and GWAS hit was calculated using the following formula in eq. 4

4. $r^2 = \frac{Cov(SNV1,SNV2)^2}{Var(SNV1) \times Var(SNV2)}$

Here, we considered signals between all three types of traits be colocalized at linkage disequilibrium > 0.5 (i.e. $LD_{caQTL\text{-}eQTL} > 0.5$, $LD_{caQTL\text{-}GWAS} > 0.5$, and $LD_{eQTL\text{-}GWAS} > 0.5$) simultaneously.

Analysis of population differentiation and adaptation

We estimate population differentiation at single-nucleotide resolution as Weir and Cockerham's F_{ST} using *vcftools* (v0.1.16) with non-default settings (--weir-fst-pop temperate_genotypes.txt --weir-fst-pop tropical_genotypes.txt) where we randomly sampled temperate lines to have the same number of individuals in each population (*61*). Chromatin accessibility QTL SNVs were split into low and high F_{ST}

groups using the R function, *ntile*, from the package *dplyr* using non-default settings (caQTL$fst, n=2). Subpopulation-specific ACRs were identified as ACRs with FDR<0.05 from multiple-tested corrected *P*-values using *p.adjust* (method="fdr") output from a Wilcoxon rank sum test using the inverse-normalized pseudobulk chromatin accessibility counts (identical to the input used for caQTL mapping). ACRs under selection were identified by intersection (bedtools intersect) with a previously generated list of genomic regions under selection (*44*).

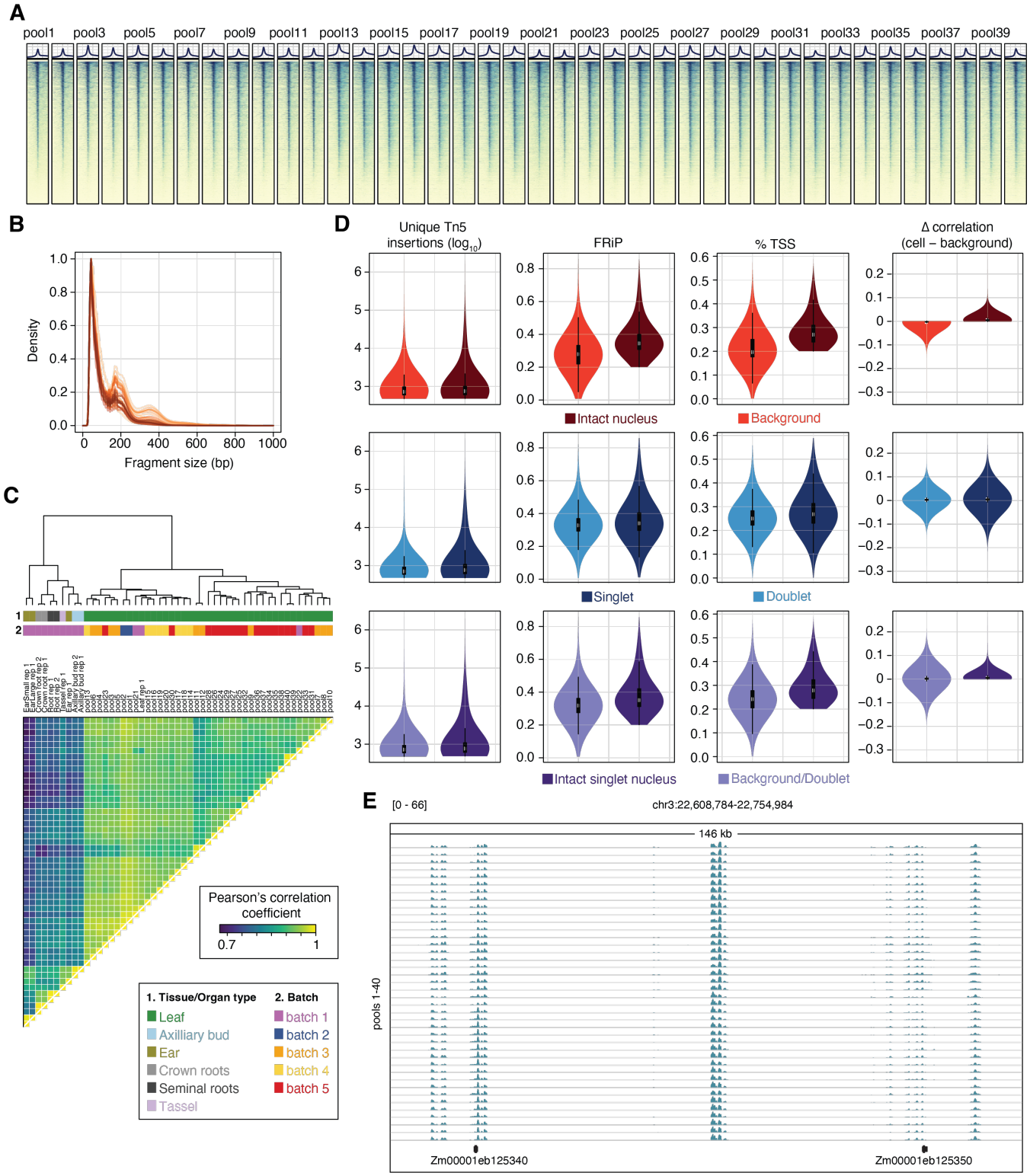

Fig. S1: Quality control analysis.
(**A**) Meta plot and heatmap of Tn5 insertions sites surrounding all ACRs across 40 genotype-pooled scATAC-seq libraries. (**B**) Fragment size distributions across all 40 genotype-pooled scATAC-seq libraries. (**C**) Correlation analysis from the union of all ACRs among the 40 seedling genotype-pooled libraries and previously published maize scATAC-seq libraries. (**D**) Barcode quality control filtering metrics. (**E**) Representative genome browser view of nuclei-aggregated normalized read depths across all 40 genotype-pooled scATAC-seq libraries. Row tracks are ordered from pool1 to pool 40.

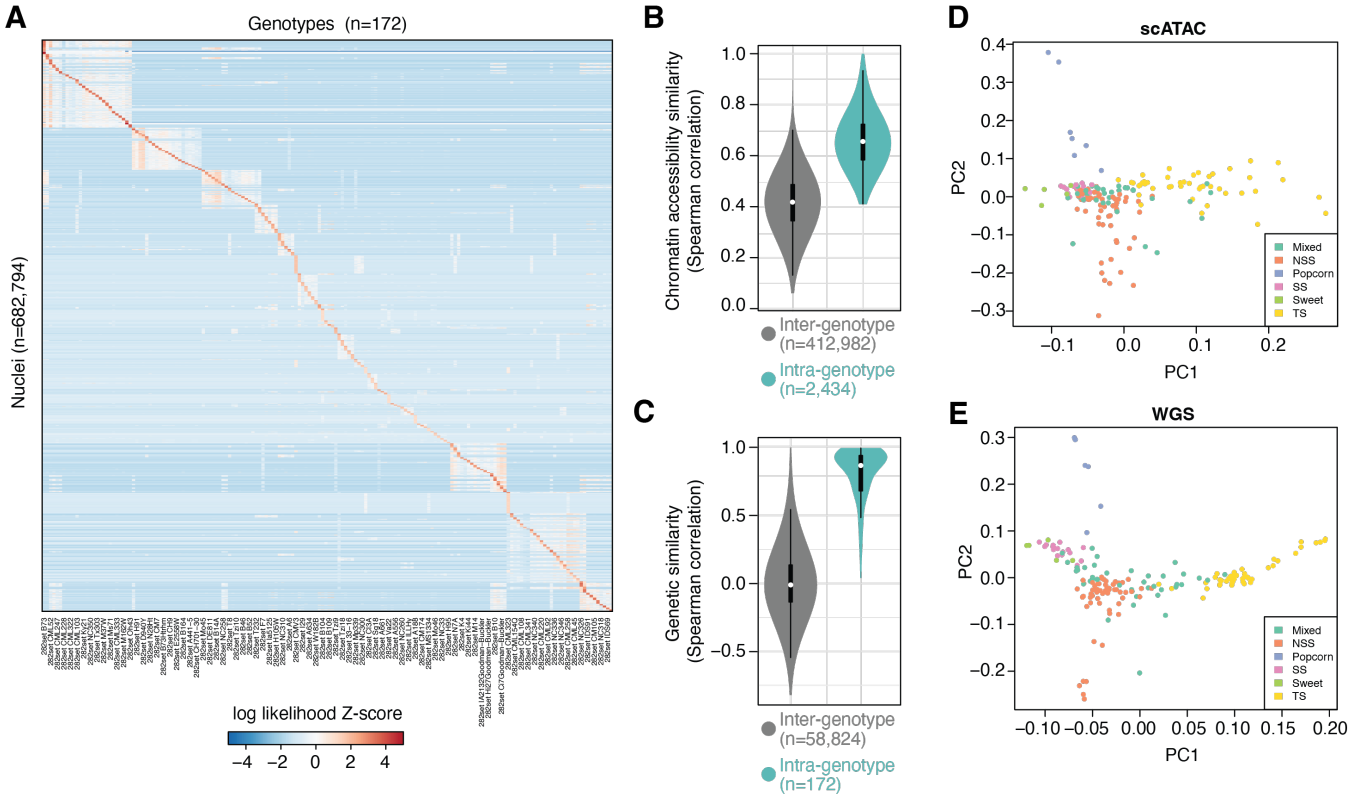

Fig. S2: Nuclei genotyping quality metrics.
(**A**) Standardized genotype assignment log likelihoods for each barcode (rows) and possible genotype (columns). (**B**) Comparison of chromatin accessibility profiles between genotypes (inter-genotyping) or between genotype replicates (intra-genotype). (**C**) Comparison of genetic similarity between genotypes or between genotype replicates. (**D**) Scatter plot of the first two PCs from Principal Components Analysis based on standardized chromatin accessibility values of nuclei-aggregates grouped by genotype. Points are colored by subpopulation assignment. (**E**) Scatter plot of the first two PCs from Principal Components Analysis based on SNVs identified within scATAC-seq nuclei-aggregates grouped by genotype. Points are colored by subpopulation assignment.

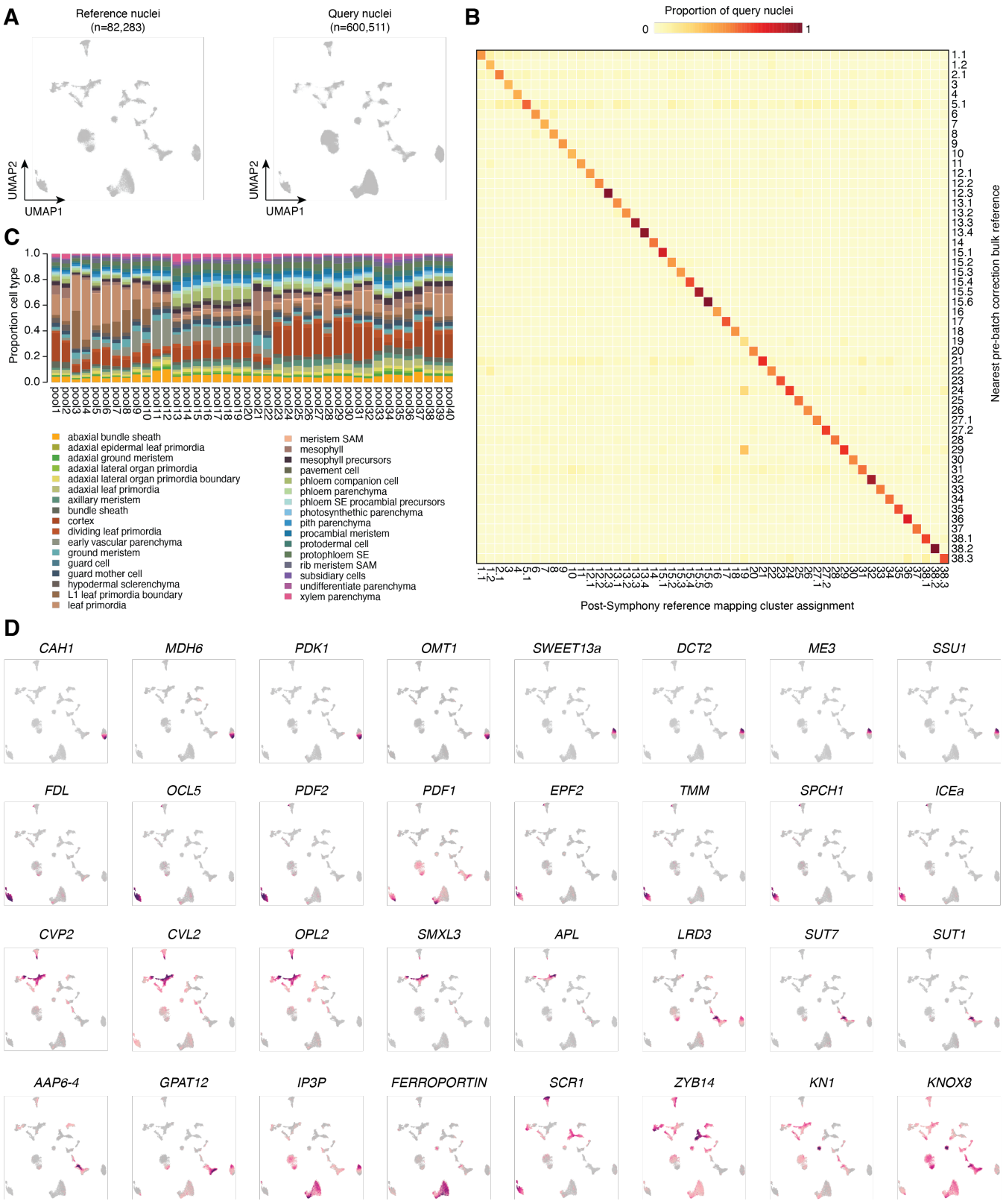

Fig. S3: Reference embedding construction.
(**A**) Comparison of UMAP embeddings of reference (left) and query (right) nuclei. (**B**) Heatmap of the proportion of query nuclei with the same nearest cluster assignment (based on Spearman's correlation to the cluster bulk) pre- and post-reference mapping. (**C**) Cell state proportions across each genotype pool. (**D**) UMAP embeddings colored by marker gene chromatin accessibility scores.

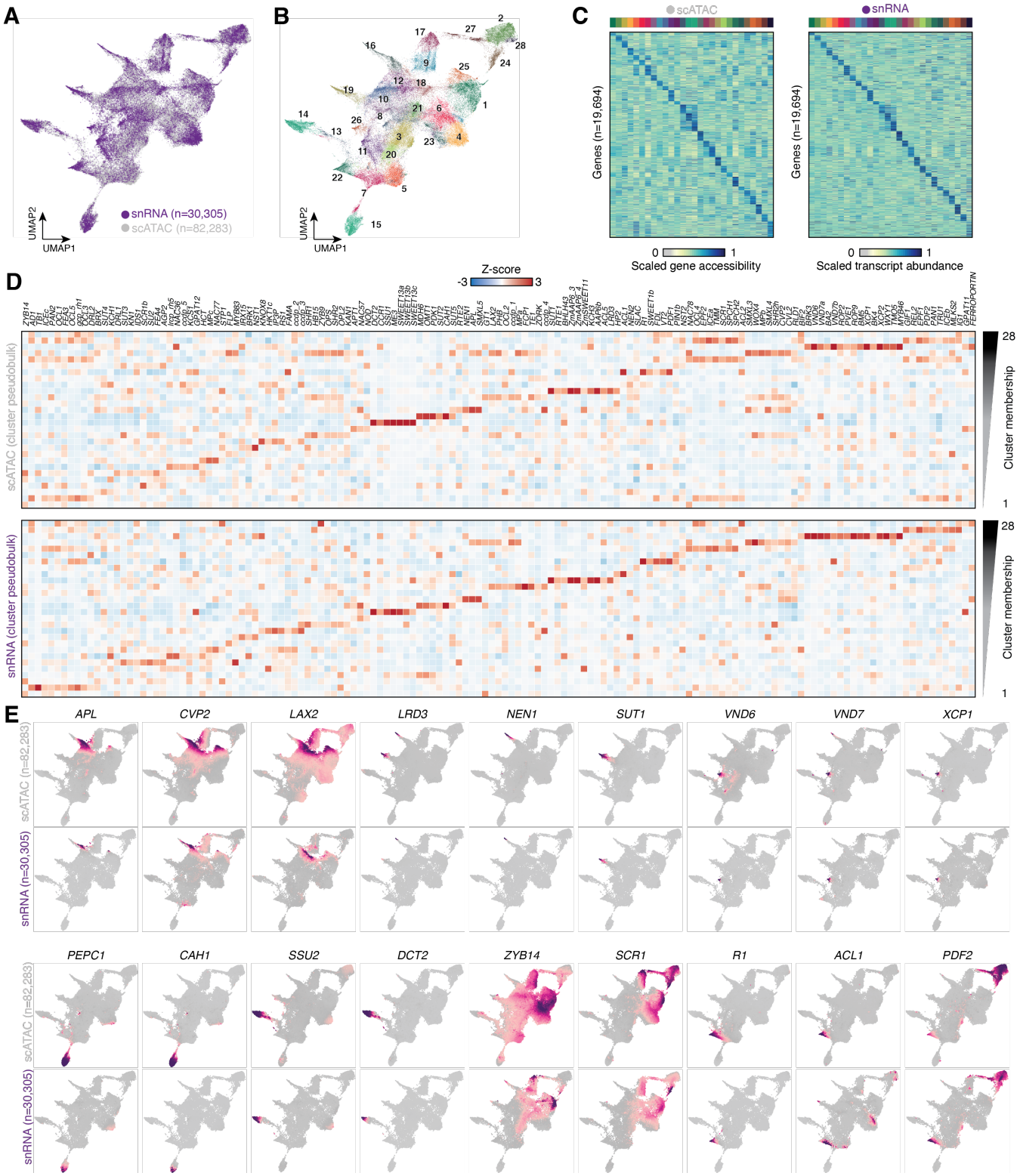

Fig. S4: Integration of seedling nuclei scATAC-seq and snRNA-seq.
(**A**) UMAP embedding of maize seedling snRNA-seq (purple) and reference scATAC-seq nuclei (grey). (**B**) UMAP embedding of leiden clusters of the integrated embedding. Nuclei are colored by cluster assignments. (**C**) Comparison of scaled gene chromatin accessibility (left) and transcript abundances (right) for cluster aggregates across 19,694 expressed genes. (**D**) Marker gene chromatin accessibility (top) and transcript abundances across clusters identified in the integrated embedding. (**E**) UMAP embedding colored by marker gene chromatin accessibility and transcript abundances for scATAC-seq and snRNA-seq nuclei, respectively.

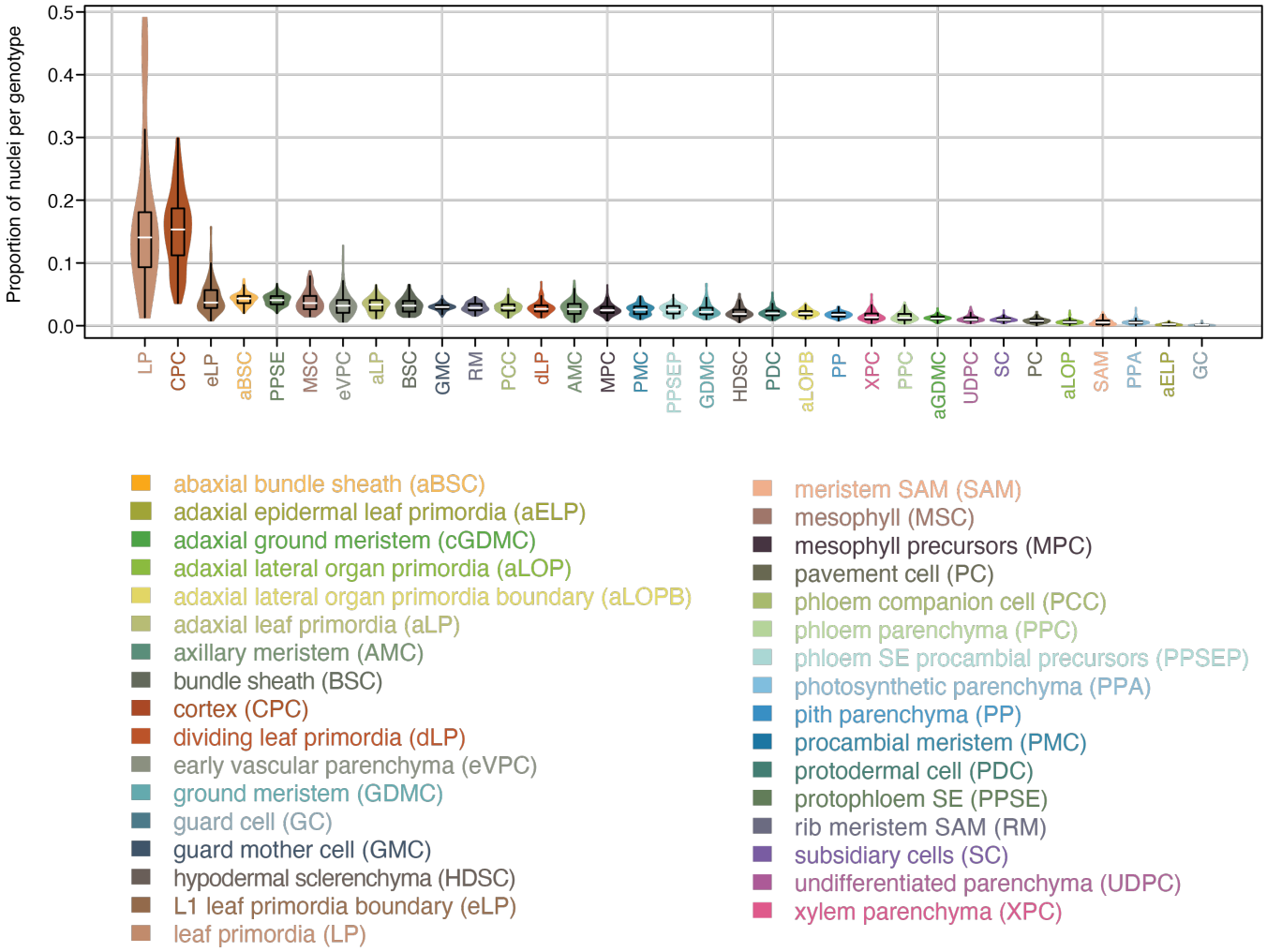

Fig. S5: Distribution of cell type proportions among genotypes. Violin plot depicting the distribution of cell type proportions across all genotypes for each individual cell state.

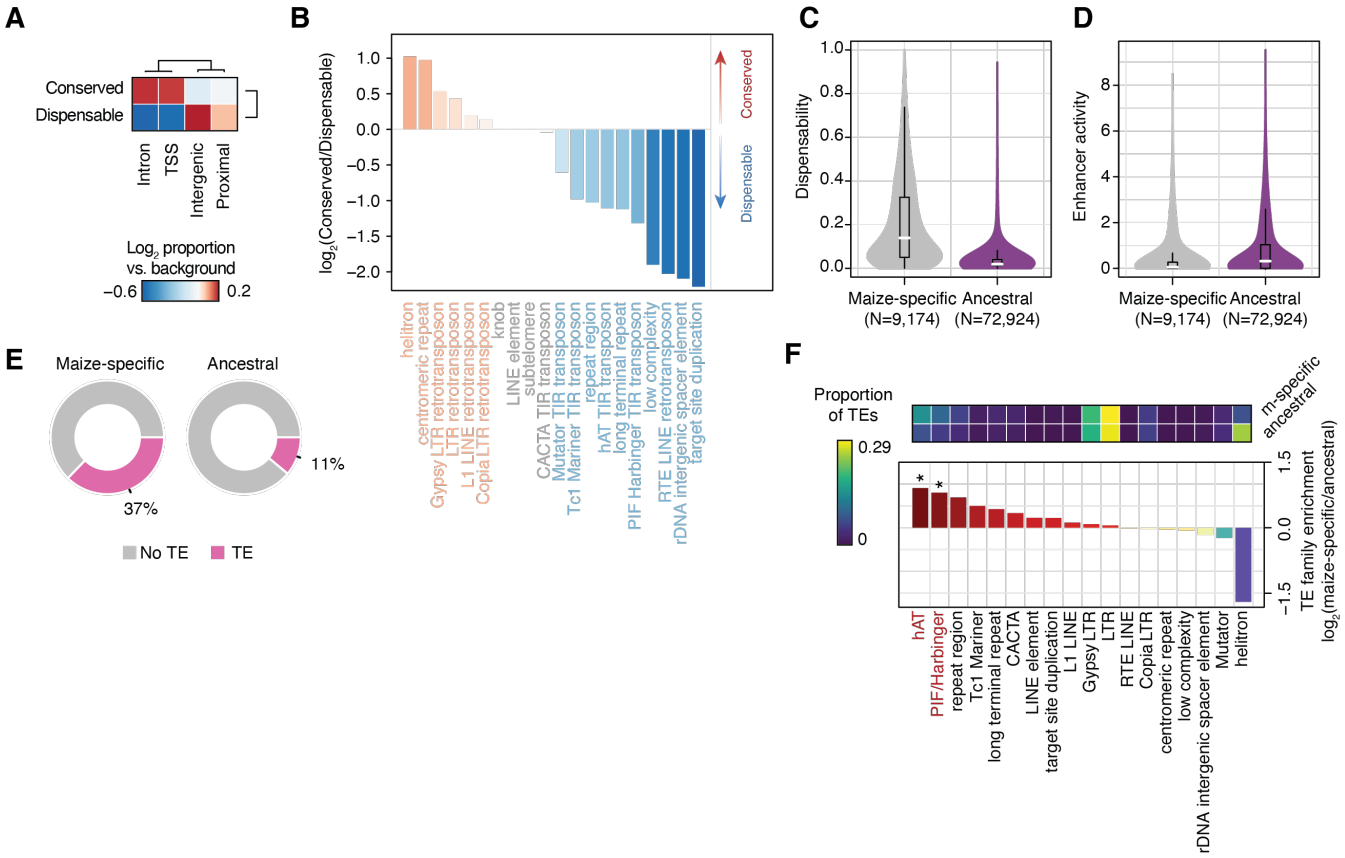

Fig. S6: Evaluation of *cis*-regulatory variation among maize inbred lines.
(**A**) Log_2 transformed proportion of conserved and dispensable ACRs compared to all ACRs across different genomic contexts. (**B**) Log_2 transformed ratio of conserved to dispensable ACRs proportions overlapping various classes of repetitive elements. (**C**) Distribution of dispensability (fraction of genotypes with WGS alignments covering less than 10% of the ACR) for maize-specific (grey) and ancestrally conserved (purple) ACRs. (**D**) Distribution of enhancer activity (log_2[RNA/DNA] from STARR-seq) for maize-specific (grey) and ancestrally conserved (purple) ACRs. (**E**) Distribution of maize-specific (left) and ancestrally conserved (right) ACRs overlapping transposable elements by at least 50% of the ACR. (**F**) Top, proportion of ACRs overlapping various classes of transposable elements split by maize-specific and ancestrally conserved. Bottom, log_2 enrichment of various transposable element families in maize-specific ACRs compared to ancestrally conserved ACRs.

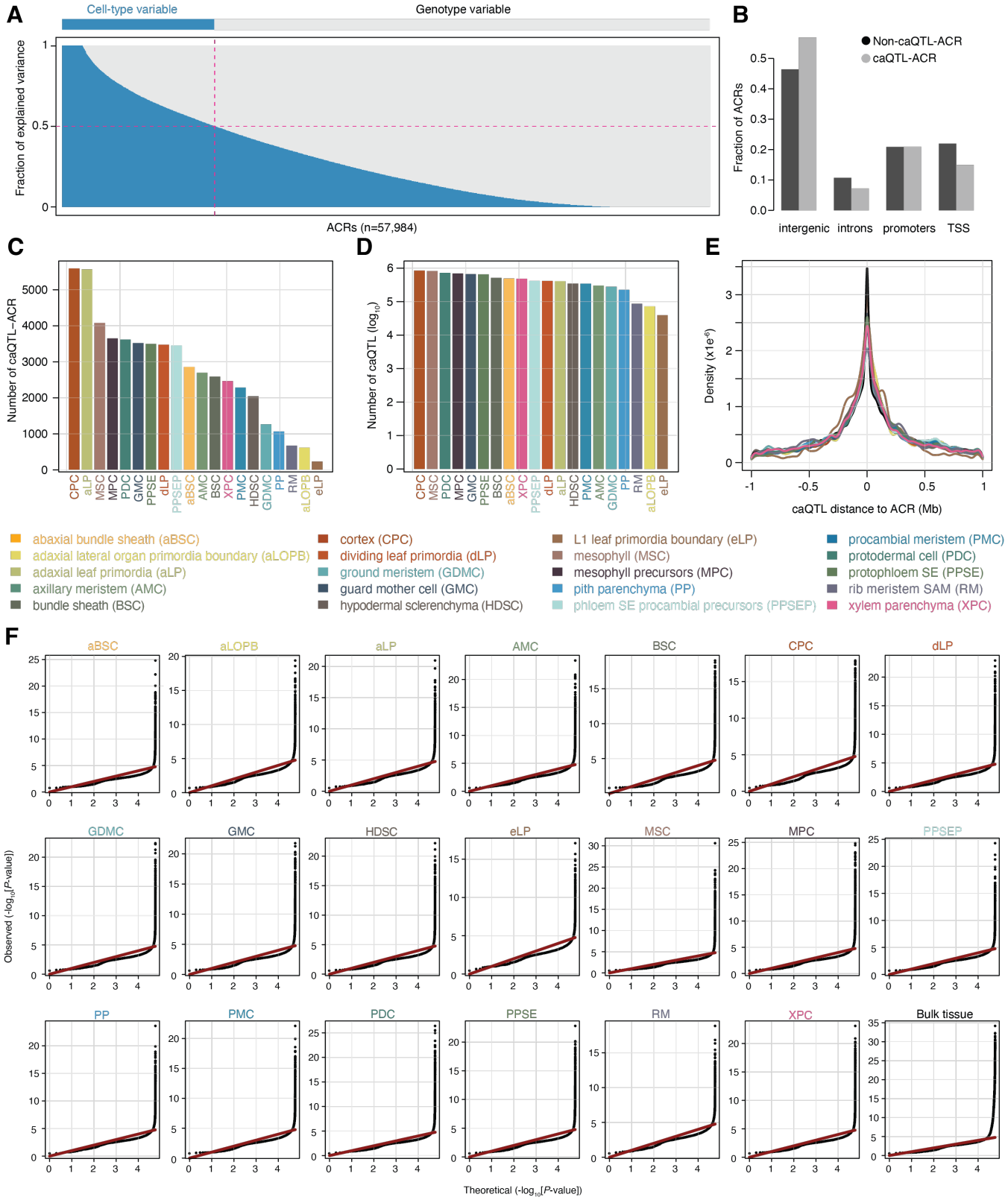

Fig. S7: Distribution of caQTL and caQTL-ACRs across cell states.
(**A**) Distribution of the relative fraction of variance explained by cell state or genotype for each ACR. (**B**) Fraction of caQTL-ACRs (grey) and non-caQTL-ACRs corresponding to various genomic contexts. (**C**) Barplot of the total counts of caQTL-ACRs by cell state. (**D**) Barplot of the $\log_{10}$ transformed counts of caQTL by cell state. (**E**) Distribution of caQTL with respect to the cognate ACR across all cell states. (**F**) Quantile-quantile plot of nominal *P*-values of the lead SNV per ACR from caQTL mapping across 20 distinct cell states and bulk tissue.

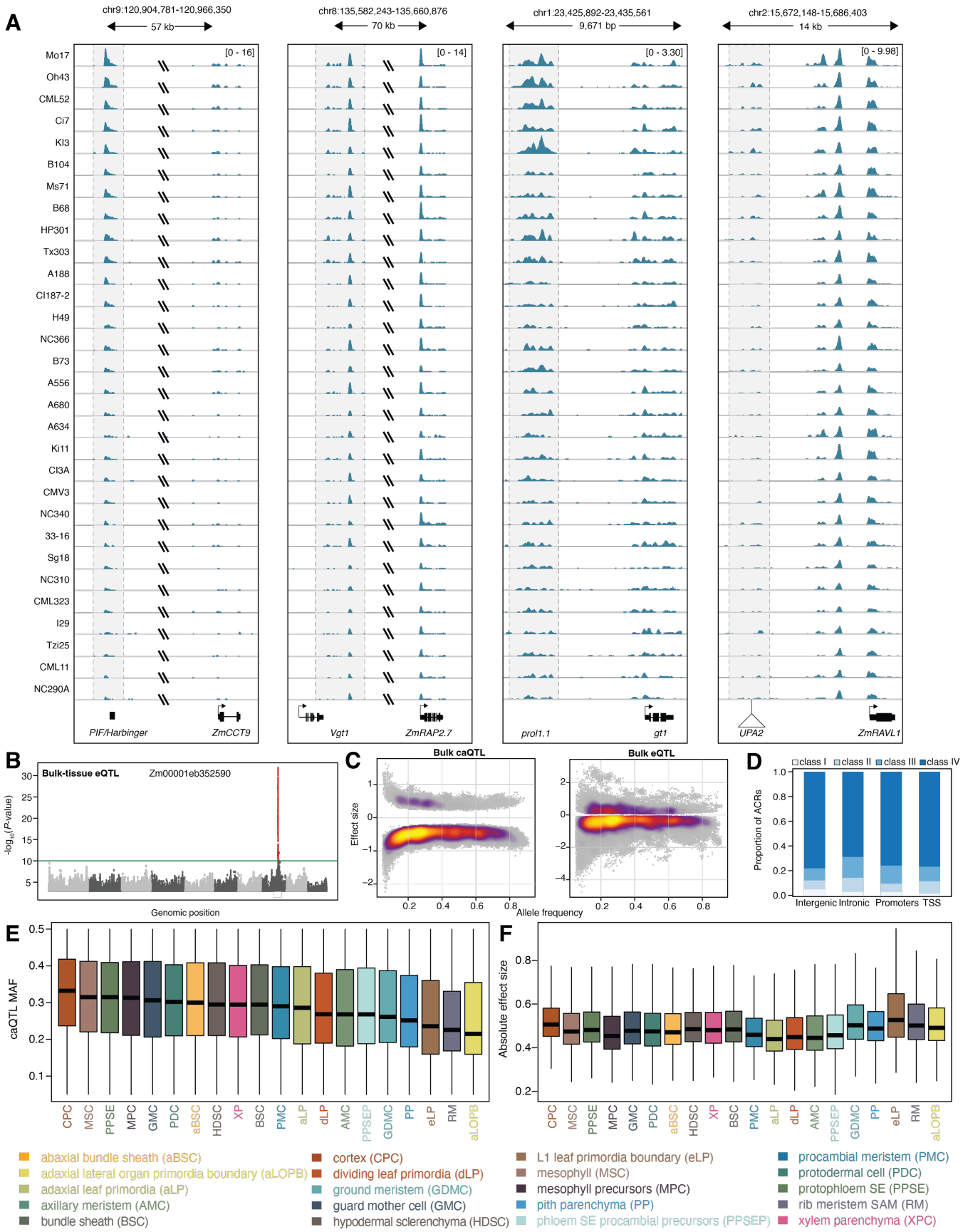
A
chr9:120,904,781-120,966,350
57 kb
chr8:135,582,243-135,660,876
70 kb
chr1:23,425,892-23,435,561
9,671 bp
chr2:15,672,148-15,686,403
14 kb
[0 - 16]
[0 - 14]
[0 - 3.30]
[0 - 9.98]
Mo17
Oh43
CML52
Ci7
Kl3
B104
Ms71
B68
HP301
Tx303
A188
Cl187-2
H49
NC366
B73
A556
A680
A634
Ki11
Cl3A
CMV3
NC340
33-16
Sg18
NC310
CML323
I29
Tzi25
CML11
NC290A
PIF/Harbinger
ZmCCT9
Vgt1
ZmRAP2.7
prol1.1
gt1
UPA2
ZmRAVL1
B
Bulk-tissue eQTL
Zm00001eb352590
-log10(P-value)
Genomic position
C
Bulk caQTL
Bulk eQTL
Effect size
Allele frequency
D
class I
class II
class III
class IV
Proportion of ACRs
Intergenic
Intronic
Promoters
TSS
E
caQTL MAF
F
Absolute effect size
CPC
MSC
PPSE
MPC
GMC
PDC
aBSC
HDSC
XP
BSC
PMC
aLP
dLP
AMC
PPSEP
GDMC
PP
eLP
RM
aLOPB
abaxial bundle sheath (aBSC)
adaxial lateral organ primordia boundary (aLOPB)
adaxial leaf primordia (aLP)
axillary meristem (AMC)
bundle sheath (BSC)
cortex (CPC)
dividing leaf primordia (dLP)
ground meristem (GDMC)
guard mother cell (GMC)
hypodermal sclerenchyma (HDSC)
L1 leaf primordia boundary (eLP)
mesophyll (MSC)
mesophyll precursors (MPC)
pith parenchyma (PP)
phloem SE procambial precursors (PPSEP)
procambial meristem (PMC)
protodermal cell (PDC)
protophloem SE (PPSE)
rib meristem SAM (RM)
xylem parenchyma (XPC)

Fig. S8: caQTL identifies known regulatory variants.
(**A**) Genome browser view of four *cis*-regulatory variants with known trait effects. (**B**) Manhattan plot of eQTL analysis for an exemplary gene. (**C**) Scatter density plot illustrating the relationship between alternate allele frequency (x-axis) and effect size (y-axis) for bulk tissue caQTL (left) and bulk tissue eQTL (right). (**D**) Proportion of caQTL/eQTL effect classifications in intergenic, intronic, promoter, and TSS-localized ACRs. (**E**) Distribution of caQTL minor allele frequency by cell state. (**F**) Distribution of absolute effect sizes of caQTL grouped by cell state.

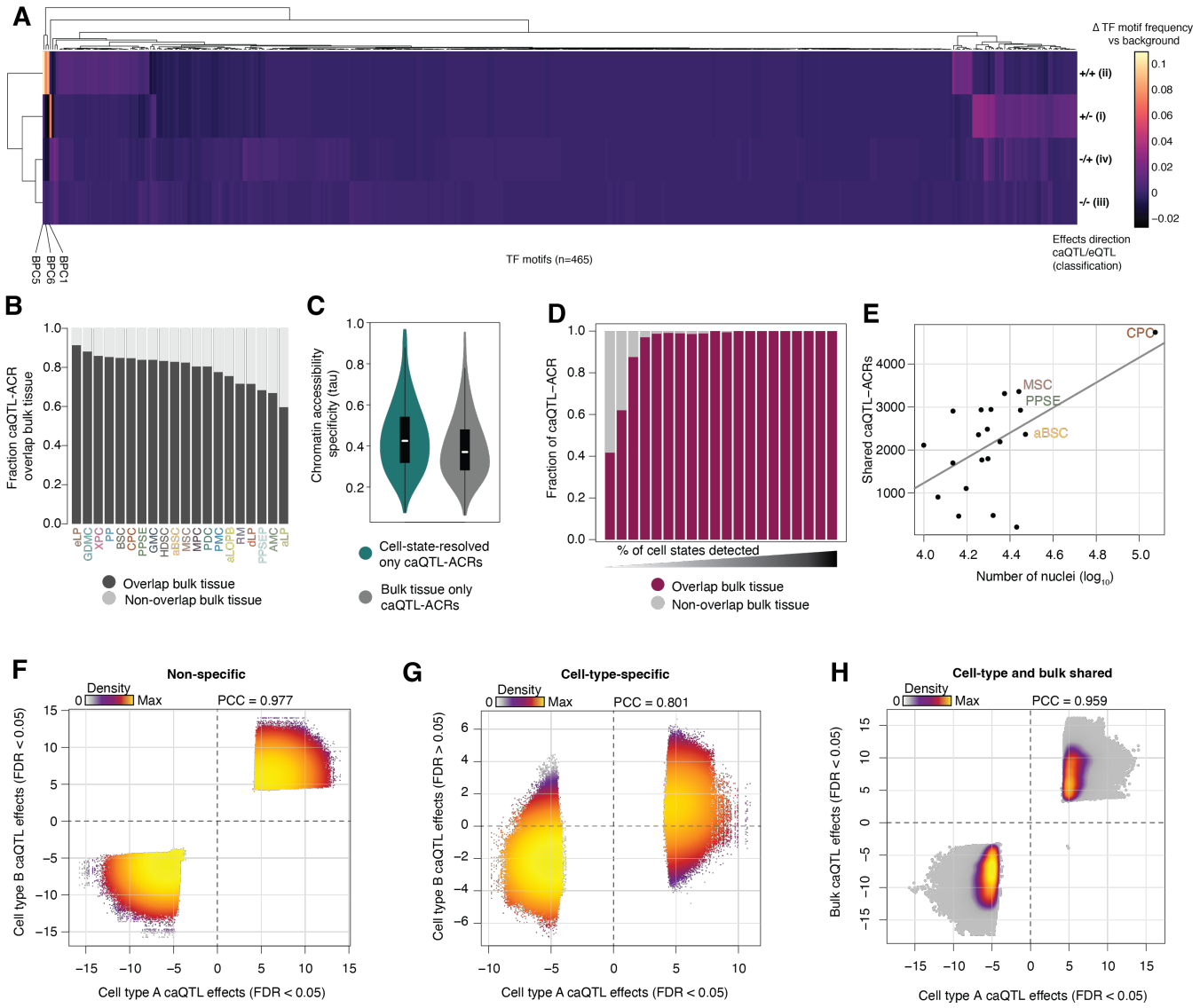

Fig. S9: Evaluation of cell-state-specific caQTL-ACRs.
(**A**) Heatmap of TFBS frequencies overlapping matched caQTL/eQTL split by effects direction classes. TFBS frequencies across all ACRs (background) was subtracted from each classification. (**B**) Fraction of cell-state-resolved caQTL-ACRs that overlap caQTL-ACRs from bulk tissue. (**C**) Distribution of ACR specificity scores (Tau) for cell-state-resolved only caQTL-ACRs and bulk tissue only caQTL-ACRs. (**D**) Fraction of cell-state resolved caQTL-ACRs that overlap bulk tissue caQTL-ACRs split by the number of cell states in which a caQTL-ACR was detected. (**E**) Scatter plot of the number of shared caQTL-ACRs between an individual cell state and bulk tissue (y-axis) against the number of nuclei identified from a given cell state (x-axis). The top four most abundant cell states are labeled. (**F**) Density scatterplot of pairwise comparison of standardized effects for the same caQTL found in at least two cell types. (**G**) Density scatterplot of pairwise comparison of caQTL standardized effects for cell-state-specific caQTL in the focal cell state (x-axis, caQTL FDR < 0.05) compared to all other cell states (y-axis, caQTL FDR > 0.05) in a pairwise fashion. (**H**) Density scatterplot of pairwise comparison of caQTL standardized effects for overlapping cell-state-resolved and bulk tissue caQTL.

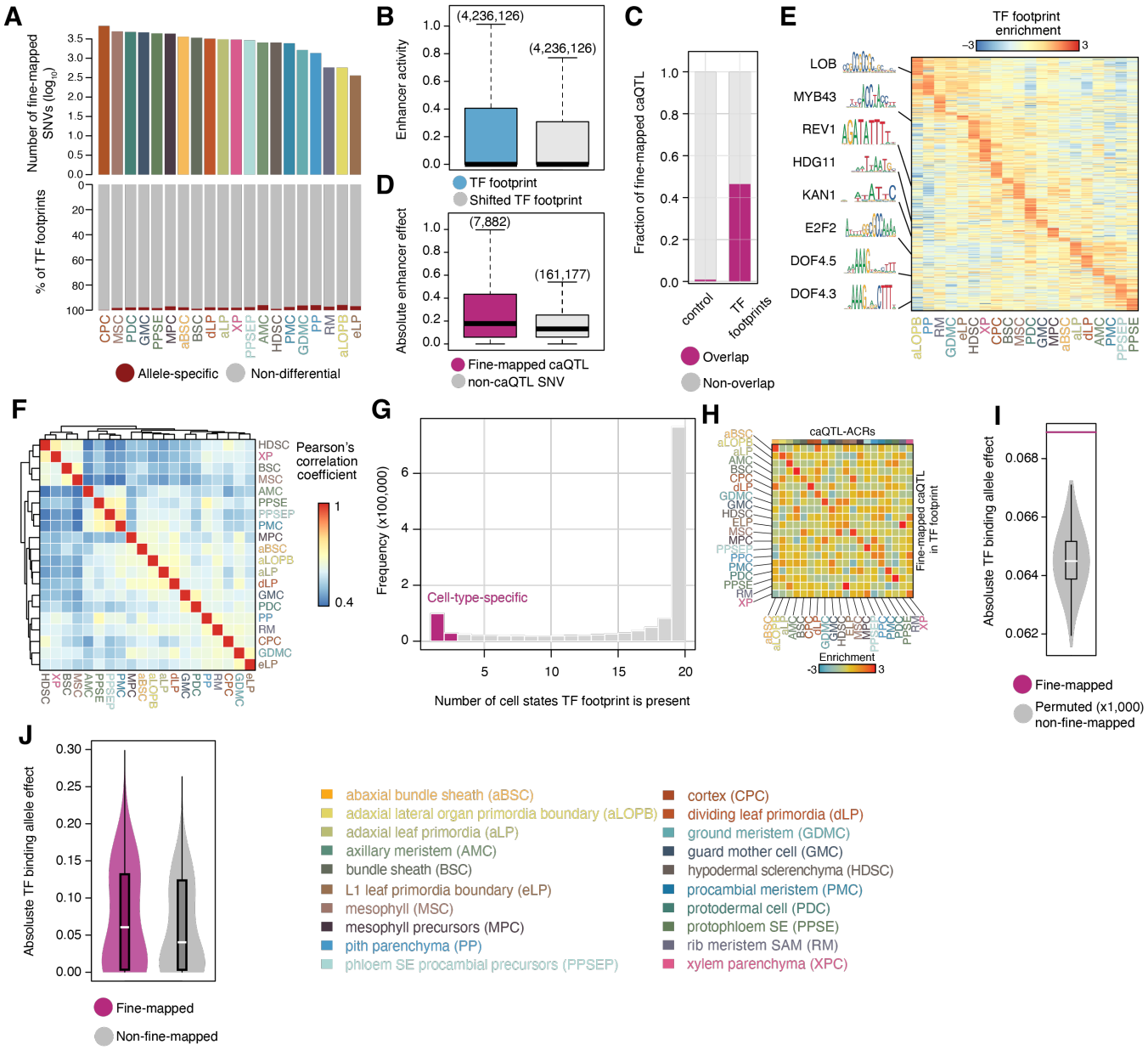

Fig. S10: Fine-mapped caQTL effect TFBSs.
(**A**) Top, barplot of the $\log_{10}$-transformed number of fine-mapped caQTL identified in each cell state. Bottom, fraction of allele-specific and non-allele-specific TF footprints across 20 cell states. (**B**) Fraction of fine-mapped caQTL overlapping TF footprints (right) or randomized control regions (left) within the same ACRs and equal number of nucleotides. (**C**) Distribution of enhancer activity overlapping TF footprints (blue) and control regions (TF footprints shifted by 50-bp). (**D**) Distribution of enhancer allelic effects for fine-mapped caQTL (left, purple), and non-caQTL variants (right, grey) located within TF footprints. (**E**) Heatmap of TF footprint enrichment for various TF motifs (rows) across 20 cell states (columns). Representative TF motifs are illustrated on the left. (**F**) Pair-wise Pearson's correlation coefficients of TF footprint enrichment between cell states. (**G**) Distribution of TF footprint frequency across cell states. Pink bars represent cell-state-specific TF footprints (TF footprints found in two or less states). (**H**) Cell-type-specific enrichment of fine-mapped caQTL embedded TF footprints with caQTL-ACRs. (**I**) Distribution of permuted absolute TF affinity allele effects from non-caQTL SNVs overlapping TFs (grey violin plot) matched to the same number of fine-mapped variants. The average TF affinity allele effect of fine-mapped variants is shown as a pink line. (**J**) Distribution of TF affinity allele effects for fine-mapped caQTL (pink) and non-caQTL SNVs (grey).

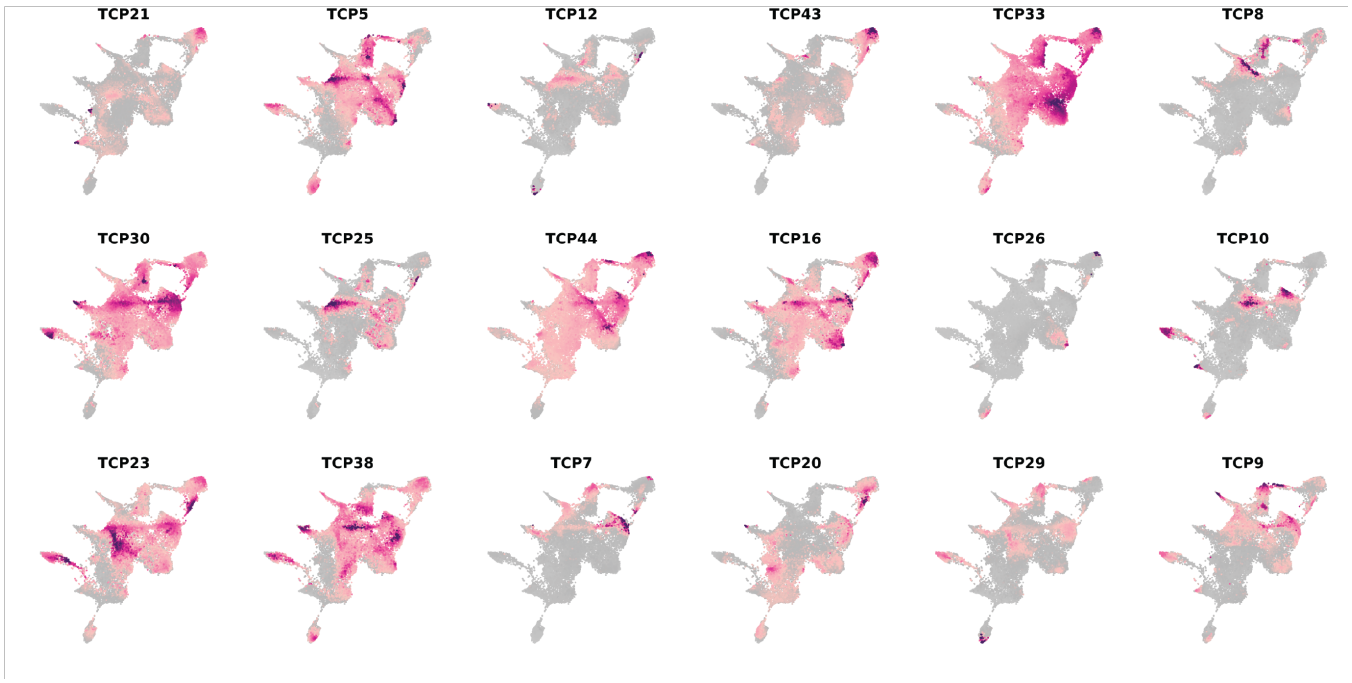

Fig. S11: Gene expression patterns for TCP transcription factors.

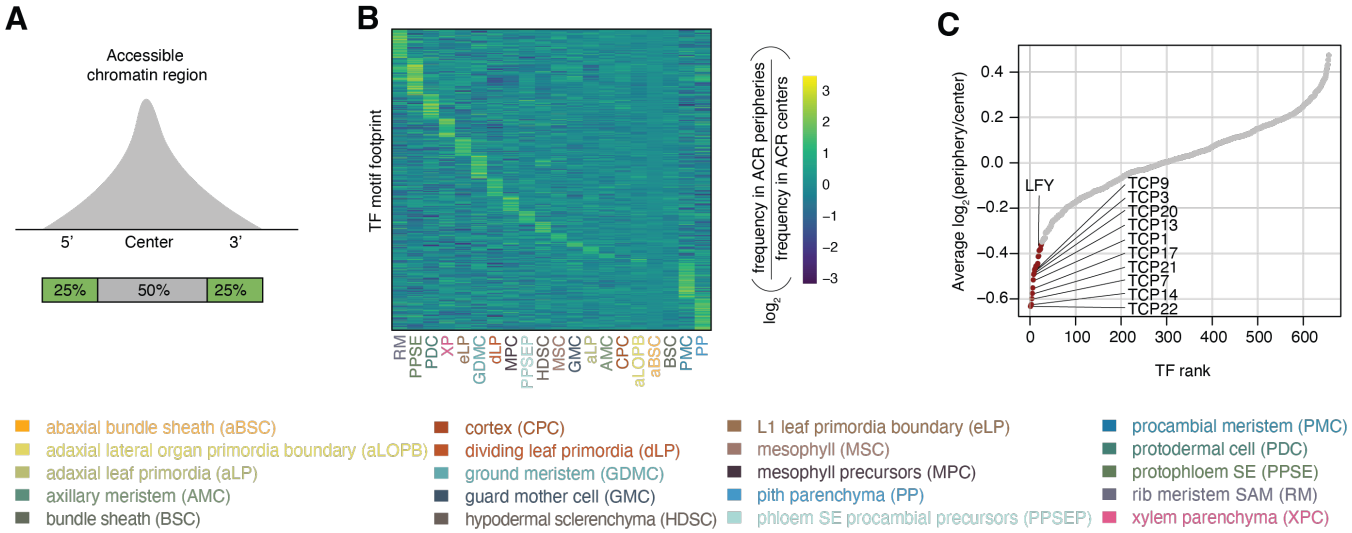

Fig. S12: Unique TFBSs are enriched in dynamic ACR boundaries: (**A**) Illustration of TF footprint assignments relative to ACR coordinates. (**B**) Heatmap of the $\log_2$ enrichment of TF footprints (rows) in ACR peripheries or ACR centers across cell states (columns). (**B**) Ranked scatter plot of the average $\log_2$ enrichment of TF footprints near ACR centers.

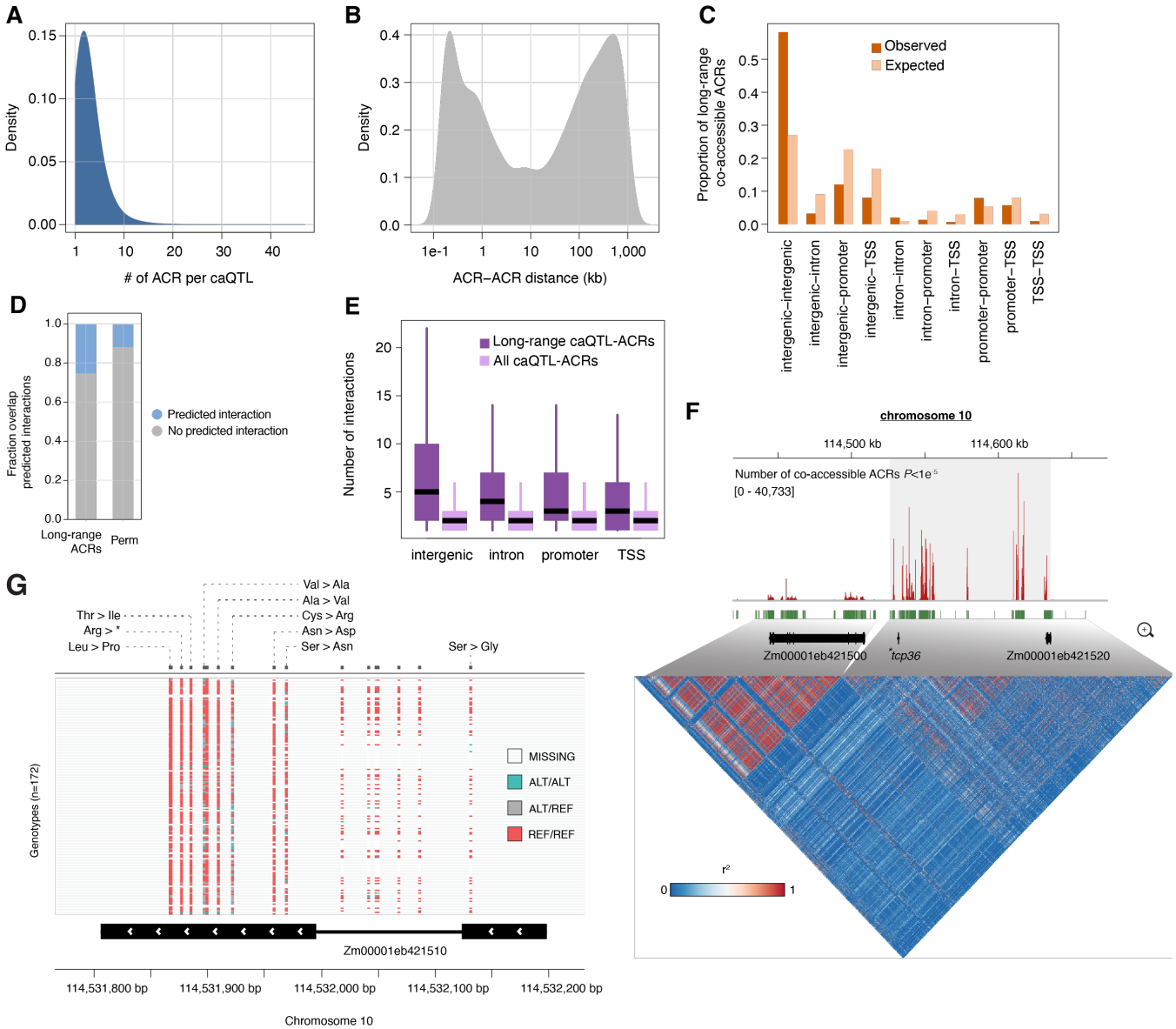

Fig. S13: Evaluation of long-range caQTL.
(**A**) Distribution of the number of ACRs associated with the same caQTL variant. (**B**) Distribution of ACR-ACR distances (kb) for long-range ACRs associated with the same caQTL. (**C**) Proportion of long-range co-accessible ACRs corresponding to various genomic contexts split by observed long-range caQTL-ACRs (observed) and all possible ACR-ACR interactions (expected). (**D**) Proportion of long-range ACRs and permuted interactions overlapping predicted interactions from the deep learning model, GenomicLinks. (**E**) Distribution of long-range interaction counts partitioned by genomic context for long-range caQTL-ACR (focused on the ACR containing the local caQTL; observed) and a random selection of the same number of ACR-ACR links derived from all caQTL-ACRs (expected). (**F**) Top: Genome-browser view of co-accessible ACR hits per SNV (red bars) in a 250-kb window surrounding *tcp36.* Green squares indicate the positions of individual SNVs. Bottom: Heatmap of linkage-disequilibrium (r^2) among all SNVs within the ~250-kb window. (**G**) Genome browser view of *tcp36* missense variants and genotype frequency across the 172 maize inbreds.

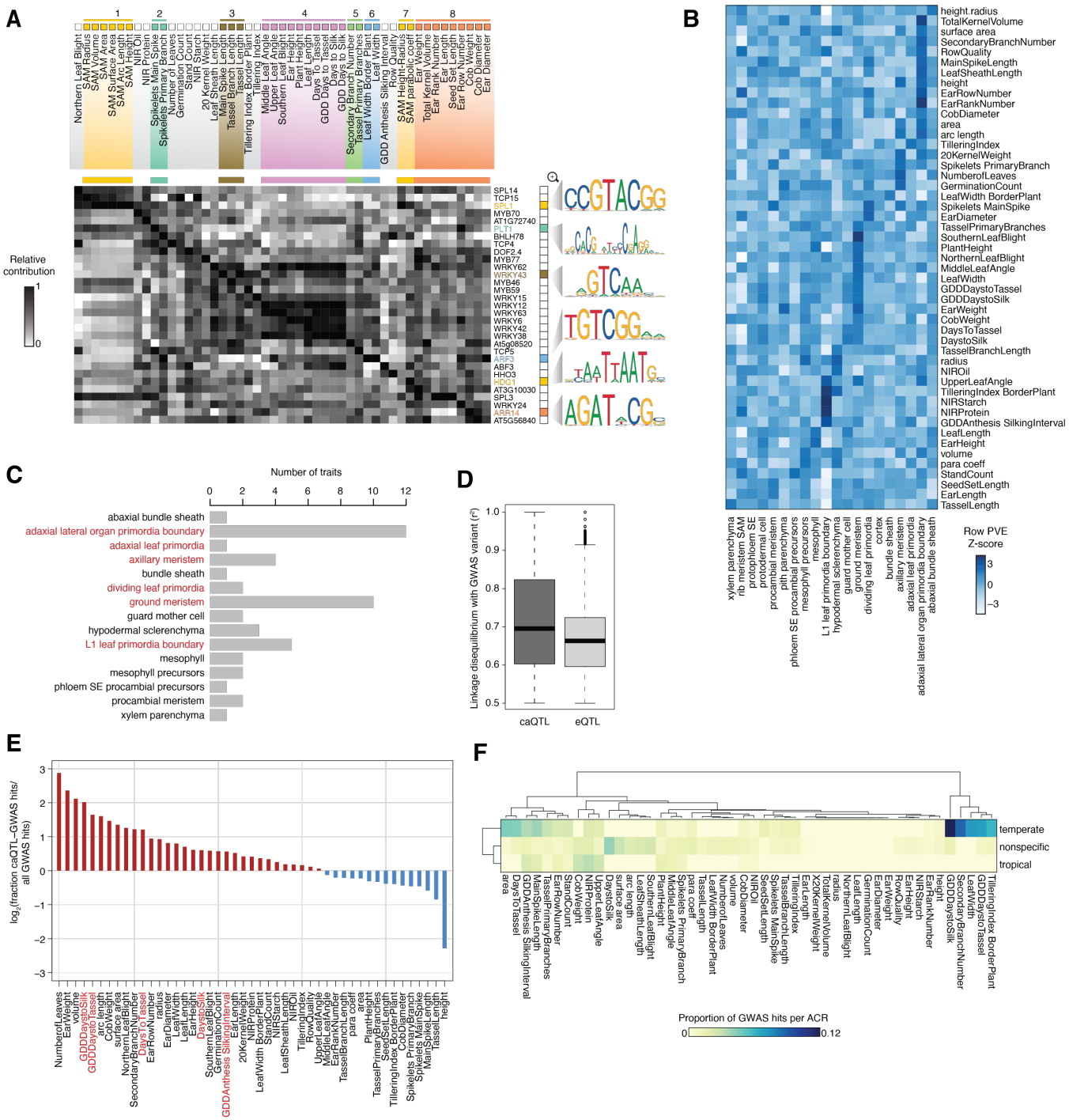

Fig. S14: Proportion of phenotypic variance explained.
(**A**) Heatmap of the top aggregated TFBS effects by phenotype. Select motifs are shown on the right. (**B**) Heatmap of standardized mean per caQTL percent of variance explained for 49 traits across 20 cell states. (**C**) Barplot of the number of traits where the cell state had the largest proportion of variance explained. (**D**) Distribution of linkage disequilibrium (r^2) between GWAS and (left) caQTL and (right) eQTL. (**E**) Log_2 frequency of GWAS hits linked to caQTL versus the overall frequency for each trait. Flowering time trait labels are denoted by red text. (**F**) Fraction of all GWAS hits mapping temperate, tropical, and non-subpopulation specific ACRs, scaled by the number of ACRs.

Captions for additional separate files:

Supplementary Data Table S1: scATAC-seq library summary statistics.

Supplementary Data Table S2: Genotype pooling strategy.

Supplementary Data Table S3: Accessible chromatin regions classifications.

Supplementary Data Table S4: Differential gene chromatin accessibility test statistics.

Supplementary Data Table S5: Single-nuclei ATAC-seq meta data.

Supplementary Data Table S6: Motif conservation summary statistics.

Supplementary Data Table S7: Fine-mapped caQTL summary statistics.

Supplementary Data Table S8: Co-accessibility *trans*-QTL mapping candidate gene list.

Supplementary Data Table S9: Z-score summary statistics from TWA.

Supplementary Data Table S10: Z-score summary statistics from CAWA.

Supplementary Data Table S11: Linked caQTL, eQTL, and GWAS hits.

Supplementary Data Table S12: Differentially accessible subpopulation ACRs.

Supplementary Data Table S13: Motif enrichment in subpopulation-specific ACRs.

Supplementary Data Table S14: Motif enrichment overlapping high and low F_{st} caQTL.

References

40. J. W. Satterlee, J. Strable, M. J. Scanlon, Plant stem-cell organization and differentiation at single-cell resolution. *P Natl Acad Sci USA* **117**, 33689-33699 (2020).

41. R. Johnston *et al.*, Transcriptomic Analyses Indicate That Maize Ligule Development Recapitulates Gene Expression Patterns That Occur during Lateral Organ Initiation. *Plant Cell* **26**, 4718-4732 (2014).

42. R. A. Kerstetter, D. Laudencia-Chingcuanco, L. G. Smith, S. Hake, Loss-of-function mutations in the maize homeobox gene, knotted1, are defective in shoot meristem maintenance. *Development* **124**, 3045-3054 (1997).

43. S. Knauer *et al.*, A high-resolution gene expression atlas links dedicated meristem genes to key architectural traits. *Genome Res* **29**, 1962-1973 (2019).

44. H. Liu *et al.*, Genomic, Transcriptomic, and Phenomic Variation Reveals the Complex Adaptation of Modern Maize Breeding. *Mol Plant* **8**, 871-884 (2015).

45. R. Tayengwa, P. Sharma Koirala, C. F. Pierce, B. E. Werner, M. M. Neff, Overexpression of AtAHL20 causes delayed flowering in Arabidopsis via repression of FT expression. *BMC Plant Biol* **20**, 559 (2020).

46. A. V. Klepikova, A. S. Kasianov, E. S. Gerasimov, M. D. Logacheva, A. A. Penin, A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. *Plant J* **88**, 1058-1070 (2016).

47. L. Y. Chen, A. Bernhardt, J. Lee, H. Hellmann, Identification of MYB56 as a Novel Substrate for CRL3 E3 Ligases. *Molecular Plant* **8**, 242-250 (2015).

48. J. Colasanti, Z. Yuan, V. Sundaresan, The indeterminate gene encodes a zinc finger protein and regulates a leaf-generated signal required for the transition to flowering in maize. *Cell* **93**, 593-603 (1998).

49. D. Zhang *et al.*, Transcription factor DIVARICATA1 positively modulates seed germination in response to salinity stress. *Plant Physiol*, (2024).

50. U. Hartmann *et al.*, Molecular cloning of SVP: a negative regulator of the floral transition in Arabidopsis. *Plant J* **21**, 351-360 (2000).

51. H. Johannesson, Y. Wang, J. Hanson, P. Engström, The homeobox gene is a potential regulator of abscisic acid responsiveness in developing seedlings. *Plant Molecular Biology* **51**, 719-729 (2003).

52. T. Kunieda *et al.*, NAC Family Proteins NARS1/NAC2 and NARS2/NAM in the Outer Integument Regulate Embryogenesis in. *Plant Cell* **20**, 2631-2642 (2008).

53. A. P. Marand, Z. Chen, A. Gallavotti, R. J. Schmitz, A cis-regulatory atlas in maize at single-cell resolution. *Cell* **184**, 3041-3055 e3021 (2021).

54. X. Zhang, A. P. Marand, H. Yan, R. J. Schmitz, scifi-ATAC-seq: massive-scale single-cell chromatin accessibility sequencing using combinatorial fluidic indexing. *Genome Biology* **25**, 90 (2024).

55. M. B. Hufford *et al.*, De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science* **373**, 655-662 (2021).

56. H. Zhang *et al.*, Fast alignment and preprocessing of chromatin profiles with Chromap. *Nat Commun* **12**, 6566 (2021).

57. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

58. Picard2019toolkit, Picard toolkit. *Broad Institute, GitHub repository*, (2019).

59. R. Bukowski *et al.*, Construction of the third-generation Zea mays haplotype map. *Gigascience* **7**, 1-12 (2018).

60. E. Garrison, G. Marth, Haplotype-based variant detection from short-read sequencing. *arXiv e-prints*. 2012.

61. P. Danecek *et al.*, The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).

62. H. Heaton *et al.*, souporcell: Robust clustering of single cell RNAseq by genotype and ambient RNA inference without reference genotypes. *bioRxiv*, 699637 (2019).

63. I. Korsunsky *et al.*, Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods* **16**, 1289-1296 (2019).

64. J. B. Kang *et al.*, Efficient and precise single-cell reference atlas mapping with Symphony. *Nature Communications* **12**, 5890 (2021).

65. K. Zhang *et al.*, A single-cell atlas of chromatin accessibility in the human genome. *Cell* **184**, 5985-6001.e5919 (2021).

66. B. Nelms, V. Walbot, Defining the developmental program leading to meiosis in maize. *Science* **364**, 52-56 (2019).

67. Y. Zhang *et al.*, Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**, R137 (2008).

68. N. Kryuchkova-Mostacci, M. Robinson-Rechavi, A benchmark of gene expression tissue-specificity metrics. *Brief Bioinform* **18**, 205-214 (2017).

69. J. D. Welch *et al.*, Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177**, 1873-1887 e1817 (2019).

70. J. Armstrong *et al.*, Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**, 246-+ (2020).

71. S. C. Stelpflug *et al.*, An Expanded Maize Gene Expression Atlas based on RNA Sequencing and its Use to Explore Root Development. *Plant Genome* **9**, (2016).

72. K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, A. Siepel, Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**, 110-121 (2010).

73. G. Hickey, B. Paten, D. Earl, D. Zerbino, D. Haussler, HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* **29**, 1341-1342 (2013).

74. X. Zheng *et al.*, A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326-3328 (2012).

75. A. Taylor-Weiner *et al.*, Scaling computational genomics to millions of individuals with GPUs. *Genome Biology* **20**, 228 (2019).

76. O. Troyanskaya *et al.*, Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520-525 (2001).

77. C. C. Chang *et al.*, Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, (2015).

78. Y. Zou, P. Carbonetto, G. Wang, M. Stephens, Fine-mapping from summary data with the "Sum of Single Effects" model. *PLoS Genet* **18**, e1010299 (2022).

79. M. Bentsen *et al.*, ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nature Communications* **11**, 4267 (2020).

80. S. G. Coetzee, G. A. Coetzee, D. J. Hazelett, motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics* **31**, 3847-3849 (2015).

81. A. P. Marand, Computational Analysis of Maize Enhancer Regulatory Elements Using ATAC-STARR-seq. *Bio-101* **14**, e4953 (2024).

82. J. Ou, S. A. Wolfe, M. H. Brodsky, L. J. Zhu, motifStack for the analysis of transcription factor binding site evolution. *Nat Methods* **15**, 8-9 (2018).

83. I. Rauluseviciute *et al.*, JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **52**, D174-D182 (2024).

84. X. Liu, M. Huang, B. Fan, E. S. Buckler, Z. Zhang, Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. *PLoS Genet* **12**, e1005767 (2016).

85. K. Zhao *et al.*, Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. *Nat Commun* **2**, 467 (2011).