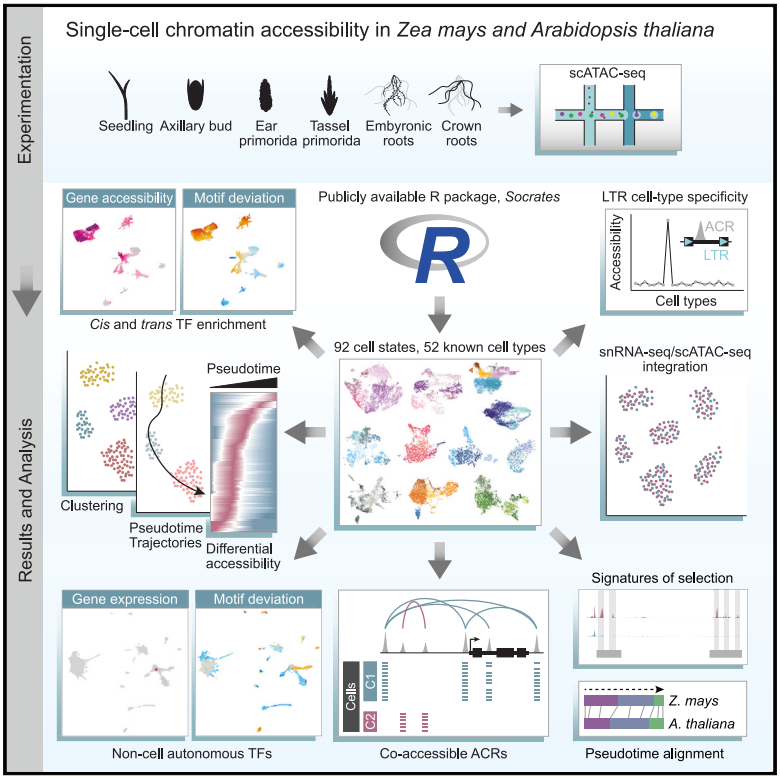
Cell

A *cis*-regulatory atlas in maize at single-cell resolution

Graphical abstract



Highlights

- Human selection and retrotransposons shaped the maize *cis*-regulatory landscape
- Analysis of 92 patterns of single-cell chromatin accessibility from six maize organs
- Transcription factors underlying chromatin interactions and non-cell autonomy
- Evolution of *cis*-regulatory dynamics during *Z. mays* and *A. thaliana* development

Authors

Alexandre P. Marand, Zongliang Chen, Andrea Gallavotti, Robert J. Schmitz

Correspondence

marand@uga.edu (A.P.M.), schmitz@uga.edu (R.J.S.)

In brief

Identification of *cis*-regulatory dynamics in the maize genome at single-cell and cell-type resolution by single-cell sequencing of assay for transposase accessible chromatin.

Marand et al., 2021, Cell *184*, 1–15
May 27, 2021
https://doi.org/10.1016/j.cell.2021.04.014

CellPress

Cell

CellPress

Resource

A *cis*-regulatory atlas in maize at single-cell resolution

Alexandre P. Marand,[1,*] Zongliang Chen,[2] Andrea Gallavotti,[2,3] and Robert J. Schmitz[1,4,*]

[1]Department of Genetics, University of Georgia, Athens, GA 30602, USA
[2]Waksman Institute, Rutgers University, Piscataway, NJ 08854, USA
[3]Department of Plant Biology, Rutgers University, New Brunswick, NJ 08901, USA
[4]Lead contact
*Correspondence: marand@uga.edu (A.P.M.), schmitz@uga.edu (R.J.S.)
https://doi.org/10.1016/j.cell.2021.04.014

SUMMARY

cis-regulatory elements (CREs) encode the genomic blueprints of spatiotemporal gene expression programs enabling highly specialized cell functions. Using single-cell genomics in six maize organs, we determined the *cis*- and *trans*-regulatory factors defining diverse cell identities and coordinating chromatin organization by profiling transcription factor (TF) combinatorics, identifying TFs with non-cell-autonomous activity, and uncovering TFs underlying higher-order chromatin interactions. Cell-type-specific CREs were enriched for enhancer activity and within unmethylated long terminal repeat retrotransposons. Moreover, we found cell-type-specific CREs are hotspots for phenotype-associated genetic variants and were targeted by selection during modern maize breeding, highlighting the biological implications of this CRE atlas. Through comparison of maize and *Arabidopsis thaliana* developmental trajectories, we identified TFs and CREs with conserved and divergent chromatin dynamics, showcasing extensive evolution of gene regulatory networks. In addition to this rich dataset, we developed single-cell analysis software, *Socrates*, which can be used to understand *cis*-regulatory variation in any species.

INTRODUCTION

The blueprints for development, response to environment, and basic function in eukaryotic cells are encoded by *cis*-regulatory elements (CREs) (Andersson and Sandelin, 2020; Marand et al., 2017). CREs contain clusters of DNA binding sites recognized by sequence-specific transcription factors (TFs) that cooperatively recruit transcriptional regulators (Gerstein et al., 2012; Ravasi et al., 2010). CRE activity is influenced by nucleosome occupancy; most TFs require nucleosome-depleted accessible chromatin to bind their target sequences (Minnoye et al., 2021). Transcriptional outcomes are dictated by interactions between core promoters and specific TFs and secondary proteins assembled at CREs. Distinct TF expression and chromatin accessibility patterns establish the gene expression programs of discrete cell types. Detailed maps of CREs and TFs in diverse cell types are essential for understanding cell function and driving innovation in biotechnologies, such as *in vivo* reprogramming of cells, cell-type-specific transgenesis, and induction of phenotypic variation via genome editing. Despite this importance, a comprehensive atlas of the CREs and TFs underlining cell identity and differentiation has yet to be realized in any plant species.

Efforts to assay TFs and CREs in plants, in contrast to mammalian models, have been limited by technical restraints imposed by the cell wall and an inability to culture cell lines. Past studies employed procedures such as isolation of nuclei tagged in specific cell types and fluorescence-activated cell sorting of GFP-tagged markers to querying plant cell types (Birnbaum et al., 2003; Brady et al., 2007; Deal and Henikoff, 2011). These methods require transgenesis and prior information regarding cell-type specificity for purification that occlude unbiased discovery. Epigenomic profiling of individual cells in plants has been limited to proof-of-principle studies in the roots of *Arabidopsis thaliana* (Dorrity et al., 2020; Farmer et al., 2020).

Systematic characterization of CREs has proven challenging, because CREs can regulate target genes through long-range chromatin interactions, spanning tens to hundreds of kilobases. In metazoan genomes, a majority of chromatin interactions are mediated by CCCTC-binding factor (CTCF) (Phillips and Corces, 2009). Plant genomes also exhibit higher-order chromatin architecture, but studied plant lineages lack an orthologous factor to CTCF, suggesting distinct molecular rules for establishing chromatin architecture in plants (Heger et al., 2012).

In addition to a fundamental role in gene regulation, mounting evidence points to genetic variation of CREs as a major source of phenotypic novelty, including disease and evolutionary divergence (Rebeiz and Tsiantis, 2017; Villar et al., 2015). A substantial proportion of phenotype-associated genetic variants have been attributed to CREs with cell-type- and development-specific activity (Hekselman and Yeger-Lotem, 2020). The link between CREs and phenotype highlights the importance for

Please cite this article in press as: Marand et al., A *cis*-regulatory atlas in maize at single-cell resolution, Cell (2021), https://doi.org/10.1016/j.cell.2021.04.014

understanding the elusive origins of cell-type- and development-specific regulatory circuitries.

Here, we describe a *cis*-regulatory atlas at single-cell resolution in the genetic model and crop species *Zea mays*. We measured chromatin accessibility and nuclear gene expression in 72,090 nuclei across six maize organs. We define the *cis*-regulatory logic underlying cell identity and detail applications of single-cell sequencing of assay for transposase accessible chromatin (scATAC-seq) to reveal TFs coordinating chromatin interactions, identify non-cell-autonomous TFs, and implicate CREs with enhancer activity and interactive capacity as substantial sources of trait variation. Through an evolutionary lens, we uncover decayed long terminal repeat (LTR) retrotransposons as contributors toward cell-type-specific circuitry, present cell-type-specific CREs underlying alleles targeted by modern breeding, and evaluate the evolutionary impacts of *cis*-regulatory variation on cellular differentiation between two highly diverged angiosperms. Finally, we present the R package "*Socrates*," a unified framework for scATAC-seq preprocessing, normalization, and downstream analysis.

RESULTS

Assembly of a *cis*-regulatory atlas in maize

To comprehensively assess *cis*-regulatory variation among cell types in a major crop, we isolated nuclei using fluorescence-activated nuclei sorting and generated single-cell chromatin accessibility profiles using scATAC-seq from six major *Z. mays* L. cultivar B73 organs (four out of six organs were biologically replicated), including axillary buds (2), staminate (1) and pistillate inflorescence (1), whole seedling (2), embryonic root tips (2), and post-embryonic crown roots (2) (Figures 1A, S1A, and S1B; Table S1). Analysis of several metrics, including biological replicates, comparison with previous bulk ATAC-seq data, transcription start site (TSS) enrichment, fragment size distributions, and genotyping mixing, was reflective of high-quality scATAC-seq data (Figures S1C–S1L; STAR Methods). In total, we identified 56,575 nuclei with an average of 31,660 unique Tn5 integrations (Table S1).

To reveal the genomic locations of putative CREs, we identified accessible chromatin regions (ACRs) by *in silico* sorting, resulting in 165,913 ACRs covering ~4% of the maize genome (STAR Methods). Because most tools for scATAC-seq analysis are tailored toward human and mouse genomes, we developed a flexible, species-agnostic model-based approach leveraging a quasibinomial logistic regression framework to remove unwanted sources of technical variation into a freely available R package termed *Socrates* (Figures S2A–S2D; STAR Methods). Following normalization with *Socrates*, we visualized similarity among nuclei by projecting into a reduced dimensional space using uniform manifold approximation projection (UMAP). This analysis revealed 10 major clusters resembling bulk organs that were further partitioned into 92 reproducible subclusters representing putative cell types with distinct chromatin profiles (Figures 1B–1D and S2E; STAR Methods).

Cell-type annotation and *in situ* hybridization

To annotate clusters with corresponding cell types, we integrated chromatin accessibility on a per-gene basis as a proxy for gene expression (Spearman's correlation coefficient [SCC] with bulk RNA-seq = 0.54–0.58; Figure S2F; STAR Methods). We then evaluated differential chromatin accessibility among clusters for a manually curated list of 221 marker genes using myriad approaches (Figures 1E, 1F, S2G, and S2H; Table S2; STAR Methods). Differential chromatin accessibility identified 74% (28,625/38,752) of genes with significant variability between clusters and a reference set of nuclei, with known marker genes associated with significantly greater cell-type specificity relative to background gene sets (empirical: $p < 1e\text{-}4$; Figures S2G and S2I; Table S3; STAR Methods). Cluster-restricted patterns of gene accessibility were consistent with known cell-type/domain-specific expression, such as co-localized accessibility of bundle sheath-specific genes *DICARBOXYLIC ACID TRANSPORTER1* (*DCT2*) and *RIBULOSE BISPHOSPHATE CARBOXYLASE SMALL SUBUNIT2* (*SSU2*), and mesophyll-specific genes *MALATE DEHYDROGENASE6* (*MDH6*) and *PYRUVATE DEHYDOGENASE KINASE1* (*PDK1*) (Figures 1E, 1F, and S2H) (Chang et al., 2012). In total, we identified 52 cell types for 83% (76/92) of clusters, capturing most expected cell types from the profiled organs (Table S2). We hypothesize that undescribed cell states are present in these data.

To corroborate predicted cell-type annotations, we performed RNA *in situ* hybridization for a subset of differentially accessible genes with no prior evidence of cell-type specificity. In all cases (five out of five), *in situ* expression patterns matched the predicted localization based on gene accessibility (Figures 1G and S3A). Estimates of cell-type proportions were concordant with prior observations, such as bundle sheath and mesophyll cells within multiple organs (Figure S3B) (Langdale et al., 1989). Marker-agnostic gene set enrichment analysis of Gene Ontology (GO) terms exemplified known cell-type functions, such as "root hair cell development" in root epidermal initials, "regulation of stomatal closure" in subsidiary cells, and "malate transmembrane transport" in mesophyll cells (Figure S3C). Cell types were generalized by highly specific GO annotations, as most (>51%) GO terms were enriched in only a handful of cell types (five or fewer), implicating chromatin accessibility dynamics as underlying the hallmarks of cell identity (Figure S3C).

Integration of chromatin accessibility and gene expression from single nuclei

Chromatin accessibility at TSSs is a well-known prerequisite for transcription. To evaluate the relationship between nuclear transcription and chromatin accessibility, we sequenced the transcriptomes of 15,515 seedling nuclei via single-nucleus RNA-seq (snRNA-seq) and integrated with matched scATAC-seq seedling data (Figure 2A; STAR Methods). Co-embedding scATAC-seq (n = 11,882) and snRNA-seq (n = 15,515) nuclei revealed 19 clusters with similar genome-wide profiles (Figures 2B and 2C; STAR Methods). Comparison of gene variability highlighted concordant patterns of chromatin accessibility and nuclear transcription across clusters, exemplified by marker genes with recognized cell-type specificity (SCC within cell types = [0.52–0.69]; Figures 2D–2F). Analysis of aggregated cell-type profiles indicated greater variation in chromatin accessibility relative to RNA expression, suggesting chromatin structure

Please cite this article in press as: Marand et al., A *cis*-regulatory atlas in maize at single-cell resolution, Cell (2021), https://doi.org/10.1016/j.cell.2021.04.014

CellPress

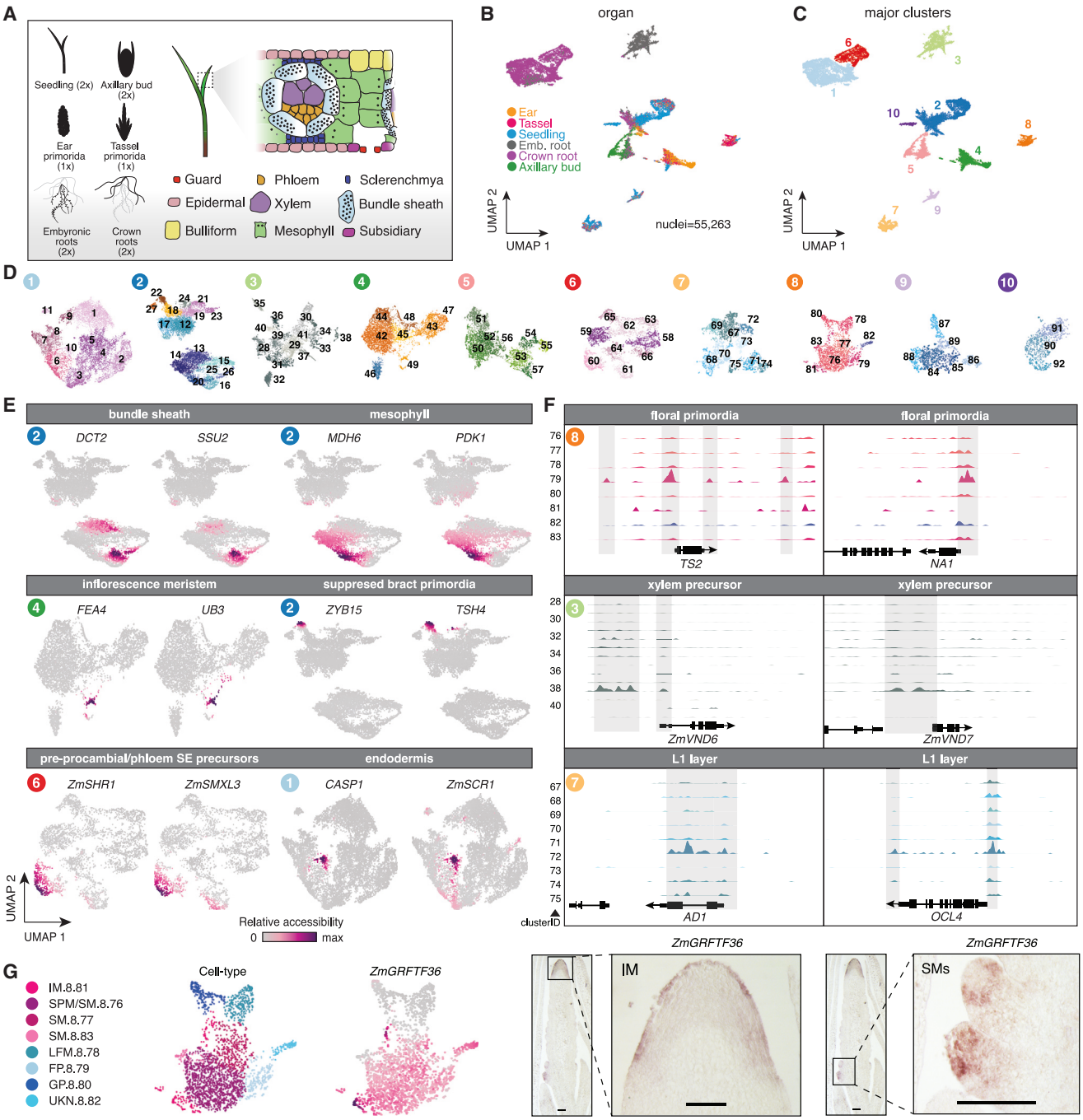


Figure 1. Profiling single-nuclei chromatin accessibility in *Zea mays*

(A) Overview of experimental samples.

(B) Two-dimensional embedding of chromatin accessibility similarity among nuclei with uniform manifold approximation and projection (UMAP). Nuclei are colored by organ identity.

(C) UMAP embedding of nuclei colored by major cluster identity.

(D) UMAP embeddings after a second round of clustering within each major cluster. Subcluster colors reflect the dominant organ of origin.

(E) Cell-type-specific gene accessibility for a subset of marker genes associated with six different cell types.

(F) Cluster-aggregated chromatin accessibility surrounding known marker genes for floral primordia, xylem precursors, and L1 epidermal cells. Numbers indicate the major cluster shown in (C). ClusterIDs refer to IDs in (D).

(G) Left: cell-type annotations and gene accessibility for *ZmGRFTF36*. Right: RNA *in situ* hybridization of *ZmGRFTF36* in staminate primordia. FP, floral primordia; GP, glume primordia; IM, inflorescence meristem; LFM, lower floral meristem; SM, spikelet meristem; SPM, spikelet pair meristem; UKN, unknown.

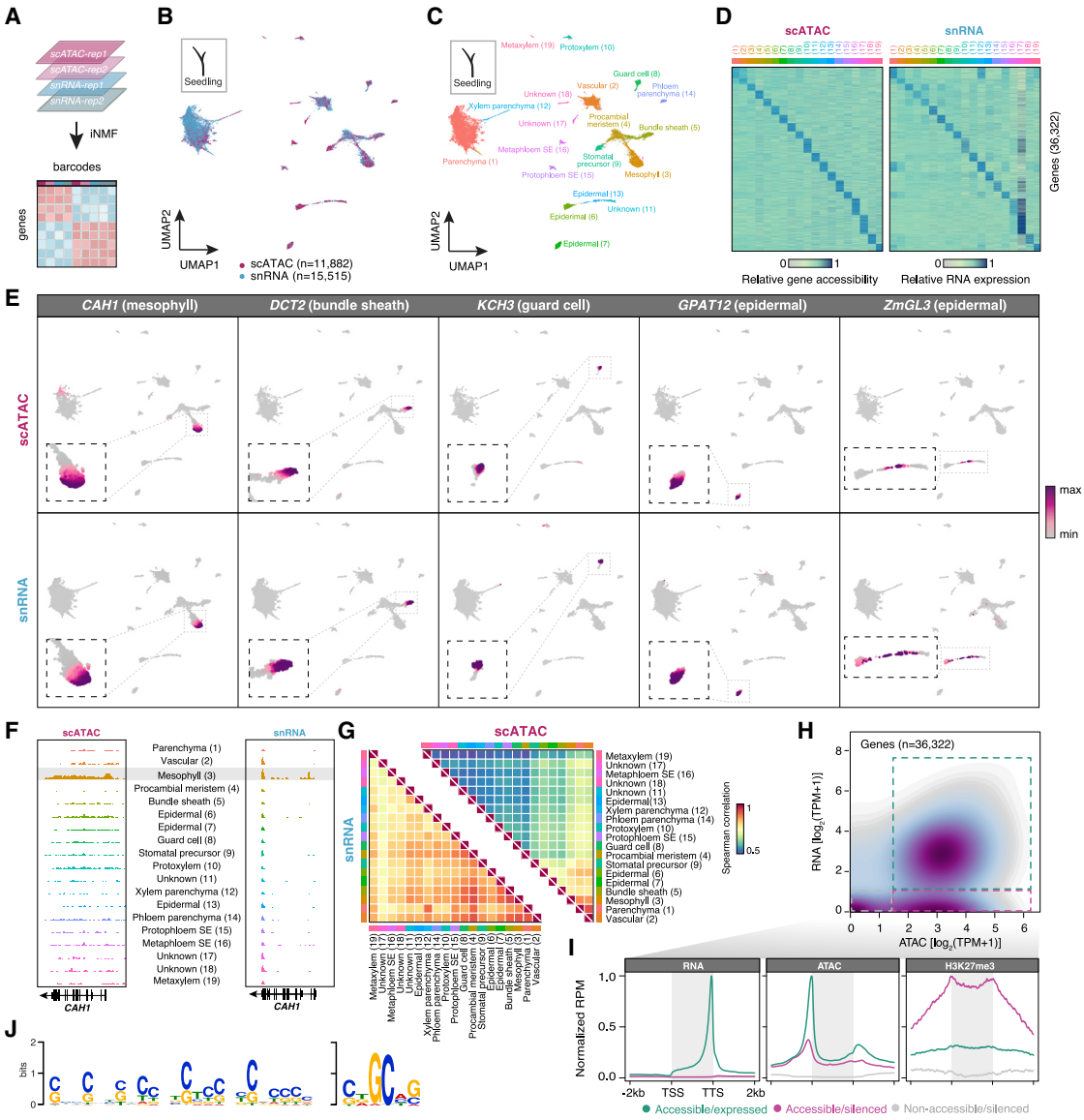


Figure 2. Gene accessibility reflects RNA expression at single-nuclei resolution

(A) Illustration of integrated non-negative matrix factorization (iNMF) integration of scATAC-seq and snRNA-seq data.

(B) UMAP co-embedding of scATAC-seq and snRNA-seq nuclei.

(C) Louvain clusters and cell-type annotations for co-embedded nuclei.

(D) Cluster-averaged gene accessibility (left) and RNA expression (right) among clusters.

(E) UMAP embeddings overlaid with gene accessibility (top) and RNA expression (bottom) for five cell-type-specific marker genes.

(F) Cluster-aggregated chromatin accessibility (left) and RNA expression (right) at the *CARBONIC ANHYDRASE1* (*CAH1*) locus.

(G) Spearman correlations between clusters based on RNA expression and gene accessibility.

(H) Density scatterplot of gene accessibility (x axis) and RNA expression (y axis) for each cluster and gene.

(I) RNA expression (left), chromatin accessibility (middle), and H3K27me3 chromatin immunoprecipitation sequencing (ChIP-seq) (right) meta-profiles (relative reads per million [RPM]) of accessible/expressed genes (n = 19,402), accessible/nonexpressed genes (n = 6,063), and nonaccessible/nonexpressed genes (n = 4,315).

(J) Two *de novo* motifs enriched in proximal ACRs of accessible/nonexpressed genes (pink, H and I).

provides additional information for dissecting cell-type heterogeneity (Figure 2G).

Despite the association between gene accessibility and expression, we observed a subset of accessible genes lacking evidence of transcription (Figure 2H). Partitioning silenced genes on the basis of chromatin accessibility revealed enrichment of H3K27me3 within and flanking accessible genes, whereas nonaccessible genes were almost entirely associated with DNA methylation (Figures 2I and S3D). Removal of H3K27me3-modified genes improved the average SCC between chromatin accessibility and

Please cite this article in press as: Marand et al., A *cis*-regulatory atlas in maize at single-cell resolution, Cell (2021), https://doi.org/10.1016/j.cell.2021.04.014

nuclear transcription (0.55–0.60), suggesting that the Polycomb repression complex (PRC) requires accessible chromatin. Considering gene accessibility includes 1 kb upstream of TSSs, we posited that proximal ACRs of accessible/silenced genes contained Polycomb response elements (PREs) capable of directing the PRC machinery. *De novo* motif analysis of 15,073 ACRs within 1 kb of accessible/silenced genes identified several enriched motifs relative to accessible/expressed genes, including a CNN repeat (E-value < 2.0e-738, 83% of ACRs, 12,858/15,073) and a CTGCAG palindromic motif (E-value < 2.4e-205, 80% of ACRs, 12,014/15,703) (Figure 2J; STAR Methods). A query with known TF binding sites revealed a significant (false discovery rate [FDR] < 4.09e-3) overlap between the CNN-repeat motif and sequences recognized by BASIC PENTACYSTEINE1 (BPC1), a BARLEY B RECOMBINANT-BASIC PENTACYSTEINE family TF associated with PREs and H3K27me3 silencing in *A. thaliana* (STAR Methods) (Xiao et al., 2017). Extending the analysis to accessible/silenced genes in root and pistillate inflorescence revealed similar *de novo* BPC1-like motifs located in proximal ACRs, suggesting PRE-based silencing may be a general regulatory mechanism in maize (Figure S3E). Taken together, our data establish gene accessibility as a robust proxy for transcription and suggests that the gene-silencing activities of certain PRCs require accessible PREs.

Genomic features of CREs

To explore CREs defining cell identity, we cataloged ACRs with discrete patterns of chromatin accessibility across cell types, identifying a total of 52,520 ACRs (31%) restricted to one or a handful of clusters (Figure S3F; STAR Methods). ACRs were prominently hypomethylated relative to the surrounding regions, consistent with prior studies (Figure 3A) (Crisp et al., 2020; Oka et al., 2017, 2020). Cluster-specific ACRs were associated with significantly greater enhancer activity determined by self-transcribing active regulatory region sequencing relative to controls and non-specific ACRs (Wilcoxon rank sum test: p < 2.2e-16; Figure 3B; STAR Methods). Deconvolution of chromatin accessibility by cell type revealed an abundance of ACRs located distal to genes (>2 kb) compared to bulk experiments (Figure 3C). Distal ACRs with enhancer activity were flanked by chromatin modifications associated with active transcription, despite being >2 kb from the nearest gene (Figures S3G and S3H). Notably, 30% (22,456/73,791) of distal ACRs overlapped LTR retrotransposons, including the maize domestication locus *TEOSINTE-BRANCHED 1*-enhancer (Figures 3C and S1A). LTRs coinciding with ACRs were associated with significantly lower levels of DNA methylation and older insertion times compared to inaccessible LTRs (empirical: p < 1e-4; Figures 3D and 3E) (Stitzer et al., 2019). Furthermore, co-localization of ACRs and LTRs was associated with greater cell-type specificity (empirical: p < 1e-4; Figure 3F). Thus, LTRs have played an important evolutionary role in wiring the regulatory landscape in maize (Noshay et al., 2020; Zhao et al., 2018).

Phenotypic variance is associated with cell-type-specific CREs

Sequence variation underlying CREs contributes to disease emergence and phenotypic innovation (Rebeiz and Tsiantis, 2017; Villar et al., 2015). To query the relationship between phenotypic variance and cell-type specificity, we quantified extant genetic variation within ACRs. Cell-type-specific ACRs were associated with a lower density of polymorphisms compared to nonspecific ACRs (Figure 3G). However, genetic variants embedded within cell-type-specific ACRs were more frequently associated with phenotypic variation determined by genome-wide association studies (Figure 3H) (Wallace et al., 2014). Thus, genetic perturbation of cell-type-specific CREs may account for a substantial proportion of phenotypic variance.

In contrast to natural populations where phenotypic changes are controlled by sexual selection, germplasm used in breeding has been subjected to selection for traits valued by humans. To determine if breeding-era selection has targeted alleles underlying cell-type-specific CREs, we assessed the relative enrichment of selection signatures from chronologically sampled elite inbred maize lines across cell-type-specific ACRs (STAR Methods) (Wang et al., 2020). Of the 21 cell types with significant (FDR < 0.01) selection signature enrichment, 57% (12) corresponded to staminate and pistillate cell types (Figure 3I). For example, a single selection block encompassing two class B TFs, *ZEA MAYS MADS 29* (*ZMM29*) and *ZMM18*, exhibited inflorescence-, spikelet-, floral-meristem-, and primordia-specific ACRs at their TSSs (Figure 3J). Dissection of allele frequencies within spikelet (pair) meristem-specific ACRs revealed signatures of selection within a MADS family TF binding site in the 5′ UTR of *ZMM29* predicted to ablate TF binding, highlighting the potential of genetic variants to affect cell-type-specific gene regulation *in vivo*. Taken together, modern maize breeding has resulted in the selection of alleles underlying floral-specific CREs that confer agronomically favorable inflorescence architectures (Gage et al., 2018).

Variation in TF activities defines distinct cell identities

To establish the TF signatures underlying distinct cells, we identified TF motifs across the maize genome. ACRs were highly enriched with TF motifs relative to control (n = 165,913) and flanking regions (Figure 4A). TF motifs were strongly depleted within ACR summits, consistent with TF-bound sequences occluding Tn5 integration (Figure 4A). To define the TF combinatorics of each cell type, we assessed the relative enrichment of TF motifs within the top 2,000 differential ACRs for each cell type. A median of 43 TF motif combinations were enriched per cell type (binomial test: FDR < 0.05; Figure 4B; STAR Methods). We hypothesized that the chromatin accessibility status of TF motifs and cognate TF genes could be used to elucidate the regulatory rules governing cell states. Comparison of TF gene accessibility with global enrichment of their sequence-specific binding sites revealed strikingly similar patterns across cell types and individual nuclei, reflecting a diverse combinatorial landscape of putatively active TFs (median SCC across cell types = 0.46; Figures 4C and 4D). Assessment of enriched TFs and their cognate motifs identified known regulators of cell identity, including *WRKY* family TFs in root epidermal progenitors and trichoblasts, *G2-like1* in parenchymal mesophyll, and *AGAMOUS-like* and *SEPALLATA* TFs in floral primordia, as well as TFs with previously unrecognized roles as cell-type regulators (Table S4) (Chang et al., 2012; Gómez-Mena et al., 2005; Verweij et al., 2016). To determine the

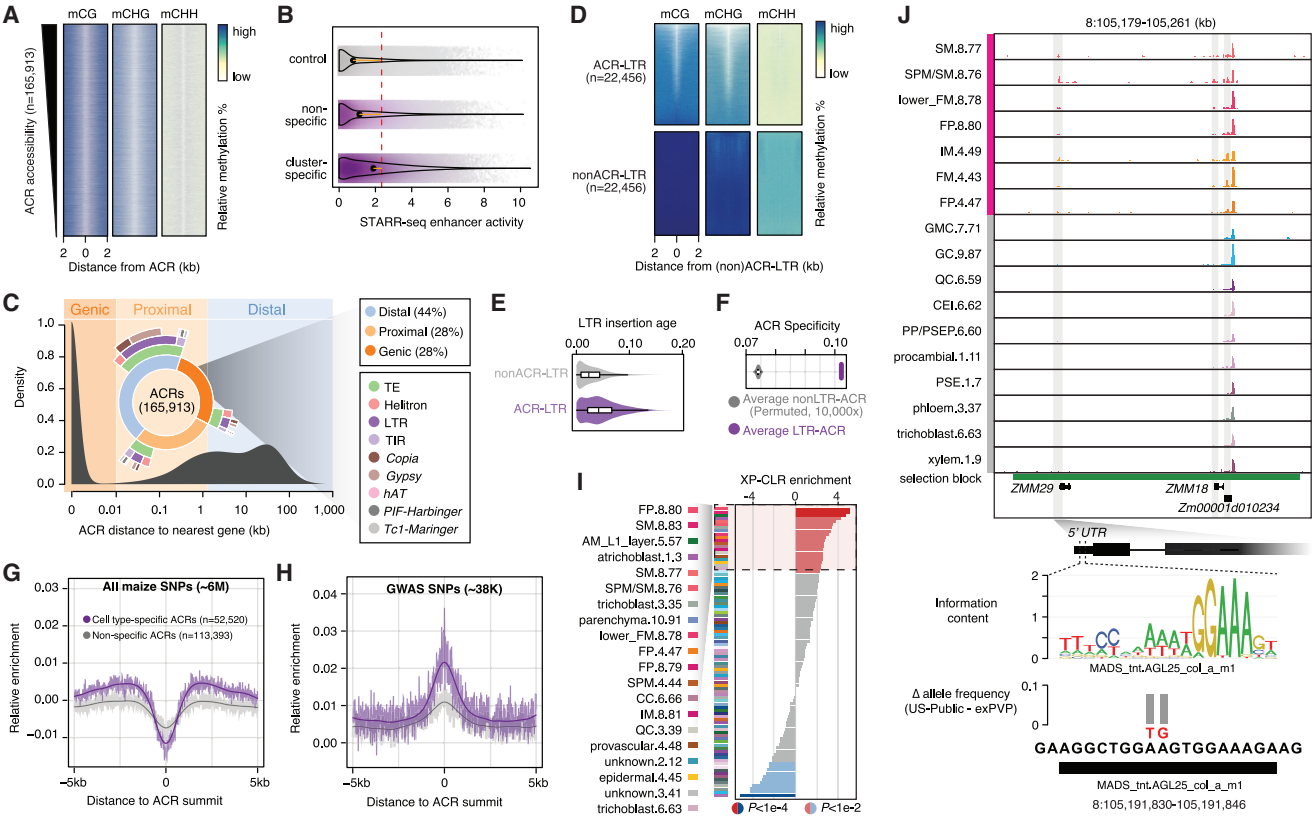


Figure 3. Characterization of ACRs

(A) Relative DNA methylation levels of 2-kb regions flanking ACRs.

(B) Distribution of enhancer activity for control regions (n = 165,913), nonspecific (n = 113,393), and cluster-specific ACRs (n = 52,520). The dash red line indicates the mean. Orange lines reflect differences between the group median and overall mean.

(C) Bimodal distribution of ACR distances to the nearest gene. Inset: distribution of ACRs by genomic context and transposon type.

(D) Relative DNA methylation levels 2-kb flanking ACRs overlapping with LTR retrotransposons (top) and LTRs devoid of accessible chromatin (bottom).

(E) Distribution of LTR insertion ages (Kimura two parameter) for non-ACR-LTRs and ACR-LTRs.

(F) Average specificity for distal ACRs overlapping LTRs (purple, n = 22,456) compared to the permuted (10,000×) average specificity for ACRs not overlapping LTRs (gray, n = 22,456).

(G) Relative SNP enrichment for 5-kb regions flanking cell-type-specific (purple) and nonspecific (gray) ACRs. Smoothed splines are shown as dark lines.

(H) Relative enrichment of genome-wide association study (GWAS) SNPs compared to all SNPs for 5-kb regions flanking cell-type-specific (purple) and nonspecific (gray) ACRs.

(I) Enrichment of signatures of selection (XP-CLR) in the top 2,000 cell-type-specific ACRs. The 20 most enriched cell types are denoted on the left. AM, axillary meristem; CC, companion cell; FM, floral meristem; QC, quiescent center.

(J) Cell-type-aggregate chromatin accessibility for seven floral cell types (pink rows) and 10 non-floral cell types (gray rows) at *ZMM29* and *ZMM18* loci. The magnified region illustrates signatures of selection coinciding with a MADS TF binding site within a floral-specific ACR at the 5′ UTR of *ZMM29*. CEI, cortex/endodermis initials; GC, guard cell; GMC, guard mother cell; PP/PSEP, pre-procambial/phloem sieve element precursor.

utility of TF motif signatures for discerning cell identity, we trained a neural network (NN) on TF motif enrichment underlying various cell types. The NN model achieved an overall accuracy of 0.94 and an average sensitivity and specificity of 0.93 and 0.99, respectively, indicating that patterns of TF motif enrichment are highly predictive of cell states (Figure 4E).

Past developmental studies have described mobile TFs capable of influencing the identities of neighboring cells. As a proxy for non-cell-autonomous activity, we searched for TFs with increased motif enrichment in cell types lacking expression of the cognate TF using the integrated snRNA-seq/scATAC-seq embedding. Of 279 TFs, we identified 20 with putative non-autonomous activity, including at least five TFs (*PHLOEM EARLY DOF1* [*PEAR1*], *TEOSINTE BRANCHED1/CYCLOIDEA/PROLIFERATING CELL NUCLEAR ANTIGEN FACTOR4* [*TCP4*], *TCP5*, *TCP14*, and *ETHYLENE RESPONSE FACTOR 018* [*ERF018*]) with predicted or known cell-cell mobility (Miyashima et al., 2019; Nag et al., 2009; Savaldi-Goldstein et al., 2007; Tatematsu et al., 2008). For example, *PEAR1* was recently described as a mobile *DNA BINDING WITH ONE FINGER* (*DOF*) family TF expressed in the procambium that promotes radial growth in the vasculature of *A. thaliana* (Miyashima et al., 2019). Consistent with predicted mobility, the maize *PEAR1* homolog, *ZmDOF36*, was expressed in procambial and protophloem cells, while its target motif was enriched in procambial, bundle sheath, phloem parenchyma, meta/protophloem, xylem, and epidermal

Please cite this article in press as: Marand et al., A *cis*-regulatory atlas in maize at single-cell resolution, Cell (2021), https://doi.org/10.1016/j.cell.2021.04.014

CellPress

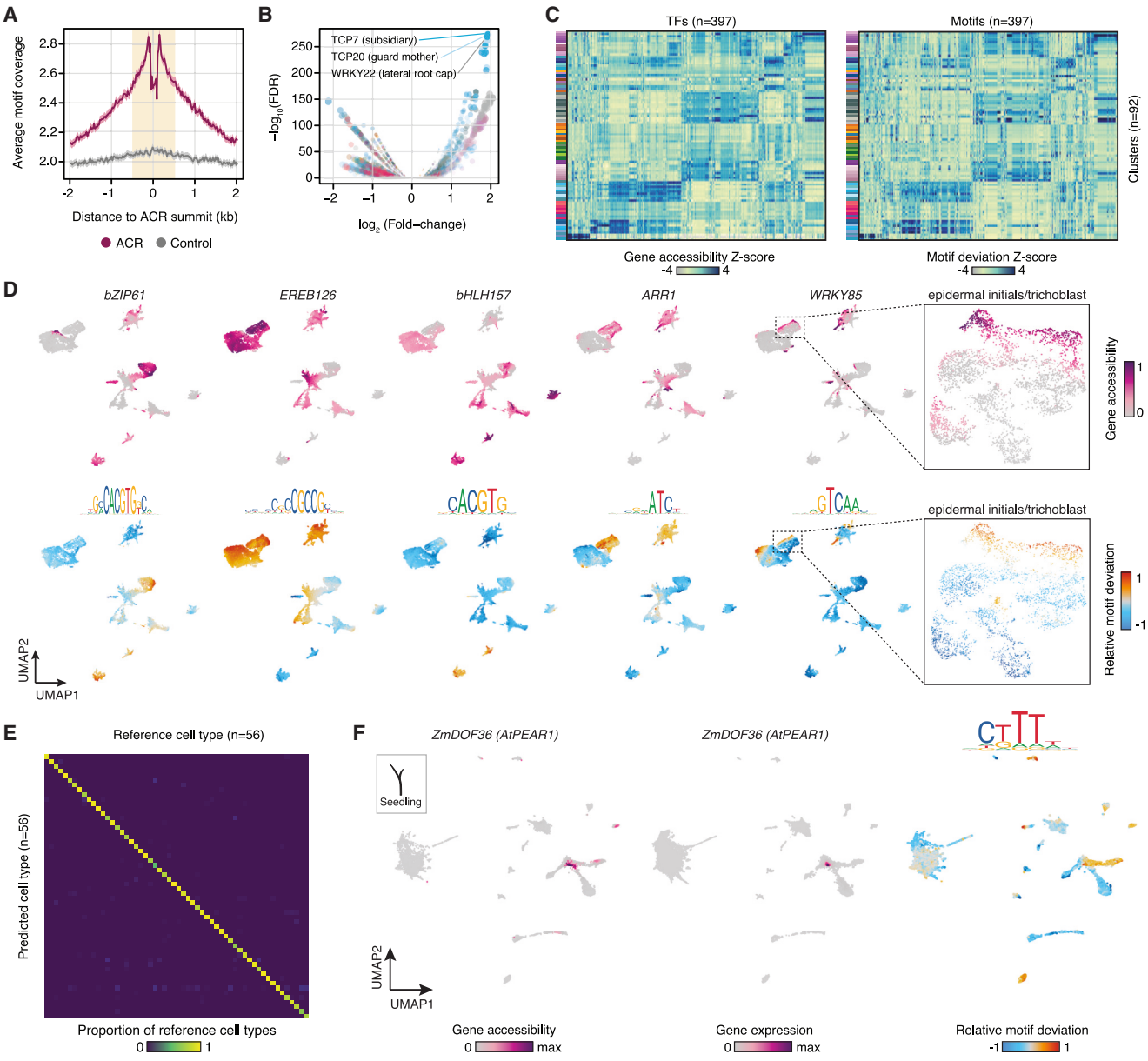


Figure 4. Combinatorial accessibility of TF genes and motifs define cell identity

(A) Average motif coverage across 4-kb windows centered on ACRs (n = 165,913) and control regions (n = 165,913). Shaded polygon, 95% confidence intervals.

(B) Enrichment of TF motifs in the top 2,000 ACRs for each cell type relative to matched constitutive ACRs.

(C) *Z* score heatmaps of TF gene accessibility (left) and motif deviation (right).

(D) Gene accessibility for five maize TFs (top) and their associated motif deviations (bottom).

(E) Predicted versus reference cell-type annotations from a NN multinomial classifier trained on combinatorial motif deviations.

(F) Gene accessibility, expression, and motif deviation for *ZmDOF36* across co-embedded seedling nuclei.

cell types (Figure 4F; Table S4). These results indicate robust inference of CRE and TF activity in single nuclei and reveal TF dynamics central to the specification of diverse cell states.

Coordinated chromatin accessibility recapitulates *in vivo* chromatin interactions

Analyses of chromatin conformation capture in plants have been instrumental in revealing chromatin organization in bulk tissues but have failed to detangle the contribution of diverse cell types (Dong et al., 2017; Liu et al., 2016, 2017; Peng et al., 2019). Leveraging recent advances for predicting chromatin architecture from single-cell data (Pliner et al., 2018), we identified 3.8 million (M) correlated patterns of chromatin accessibility between nearby ACRs (co-accessible ACRs), including known physical interactions such as *tb1*, maize *RELATED TO AP2.7* (*ZmRAP2.7*), and *BENZOXAZINLESS 1*

Please cite this article in press as: Marand et al., A *cis*-regulatory atlas in maize at single-cell resolution, Cell (2021), https://doi.org/10.1016/j.cell.2021.04.014

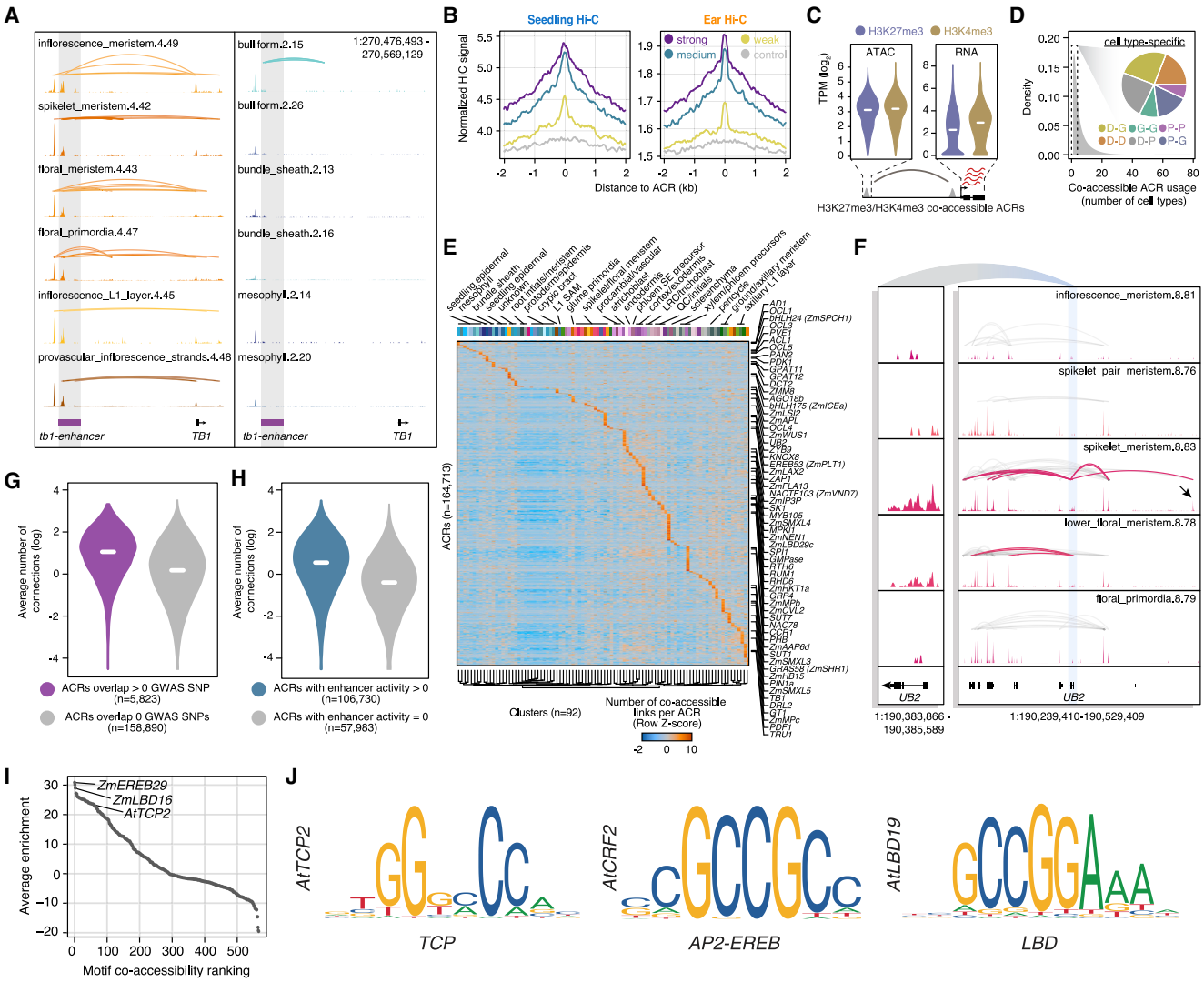


Figure 5. Co-accessible ACRs reflect *in vivo* chromatin interactions

(A) Ear and seedling co-accessible ACRs for nine cell types surrounding *TB1*. Link height reflects co-accessibility between ACRs.

(B) Average normalized Hi-C signal across 4-kb windows centered on ACRs split into terciles, each with n = 54,904 ACRs, based on the number and strength of participating co-accessible links.

(C) Chromatin accessibility (left) and RNA expression (right) distributions for H3K27me3- and H3K4me3-HiChIP chromatin interactions intersecting with co-accessible ACRs.

(D) Frequency of co-accessible ACRs across clusters. Inset: genomic context of cell-type-specific co-accessible ACRs. D, distal; P, proximal; G, genic.

(E) Co-accessible link count *Z* scores associated with each ACR within a given cell type. ACRs near marker genes are indicated on the right.

(F) Chromatin accessibility and co-accessible ACRs surrounding *UB2*, the gene underlying an inflorescence row and tassel branch number quantitative trait locus (right). The black arrow indicates a spikelet meristem-specific distal ACR. Pink, co-accessible ACRs associated with *UB2;* gray, non-*UB2*. Link height represents co-accessibility. Chromatin accessibility of *UB2* (left).

(G) Distributions of log-transformed average co-accessible ACR counts across cell types split by ACRs that do (purple) and do not overlap (gray) GWAS SNPs. White horizontal lines indicate the median. Violin plots present the entire range.

(H) Distributions of the log-transformed average co-accessible ACR counts across cell types for ACRs with (blue) and without (gray) enhancer activity.

(I) Motifs ranked by the average co-occurrence in co-accessible ACR edges relative to randomized ACR-ACR connections across cell types.

(J) Exemplary motifs enriched in reciprocal co-accessible ACRs for *TCP*, *AP2-EREB*, and *LBD* TF families.

(*BX1*) (Figures 5A and S4A–S4D; STAR Methods) (Clark et al., 2006; Salvi et al., 2007; Zheng et al., 2015). To assess the concordance of co-accessible ACRs with *in vivo* interactions, we compared co-accessible ACRs from seedling cell types with maize seedling chromatin conformation capture data, recovering 78% (3,313/4,265), 57% (37,712/65,691), and 44% (17,108/38,567) of chromatin loops from chromatin conformation capture coupled sequencing (Hi-C), H3K4me3 Hi-C chromatin immunoprecipitation (HiChIP), and H3K27me3-HiChIP experiments, respectively (Figure S4E). The integrated strength

Please cite this article in press as: Marand et al., A *cis*-regulatory atlas in maize at single-cell resolution, Cell (2021), https://doi.org/10.1016/j.cell.2021.04.014

of co-accessibility per ACR was consistent with *in vivo* chromatin interaction frequencies, indicating co-accessible ACRs are suitable proxies for chromatin loops (Figure 5B; STAR Methods).

We posited that co-accessible ACRs coinciding with H3K4me3 and H3K27me3-HiChIP chromatin loops may be associated with distinct transcriptional outcomes. To test this, we imputed nuclear RNA gene expression onto seedling scATAC-seq nuclei using the integrated embedding and compared RNA expression levels between genes associated with H3K4me3 and H3K27me3-HiChIP chromatin loops (STAR Methods). Although ACR coverages were relatively similar, genes associated with H3K27me3 co-accessible ACRs had significantly lower expression, consistent with a silencing function of gene-distal CREs flanked by H3K27me3 (empirical: $p < 1e-4$; Figure 5C) (Cai et al., 2021; Ngan et al., 2020).

Cataloguing co-accessible ACRs identified 27% (~1 M in total), 11,069 on average, that were unique to a single cell type (Figure 5D). Proximal ACRs, rather than distal or genic ACRs, were associated with a greater number of links (empirical: $p < 1e-4$; Figure S4F). Highlighting long-range "hub" interactions as key contributors toward cell identity, ACRs with cell-type-specific co-accessibility were associated with a greater number of links (empirical: $p < 1e-4$) and a greater proportion of links involving distal ACRs (empirical: $p < 1e-4$; Figures S4G and S4H). Furthermore, the interactive capacity of an ACR strongly depended on cell-type context (Figure 5E). For example, *UNBRANCHED 2* (*UB2*), a major ear row number and tassel branch number quantitative trait locus (Chuck et al., 2014), demonstrated preferential accessibility in spikelet meristems coinciding with the greatest number of *UB2* proximal-distal ACR interactions, including a cell-type-specific ACR 150 kb upstream (Figure 5F). We hypothesized that ACRs with expanded interactive capacity resemble enhancers with the potential to influence traits. Indeed, ACRs with enhancer activity and co-localization with phenotype-associated genetic variants were associated with a greater number of links (empirical: $p < 1e-4$; Figures 5G and 5H). These results highlight diverse cell-type-specific regulatory configurations among distal enhancers and their target genes and implicate highly interactive distal enhancers as major contributors toward phenotypic variation.

The structural protein, CTCF, plays an important role in mammalian genome organization and is absent in plant lineages (Heger et al., 2012). We posited that chromatin structure captured by co-accessible ACRs may be driven by distinct TFs. Motif composition of ACRs apart of co-accessible links were more similar compared to permuted links (empirical: $p < 1e-4$; Figure S4I). In addition, several TF motifs exhibited reciprocal enrichment in the edges of co-accessible ACRs (FDR < 0.05; Figure S4J; STAR Methods). We identified TCP, APETALA2/ETHYLENE-RESPONSIVE ELEMENT BINDING PROTEINS (AP2-EREBP) and LATERAL ORGAN BOUNDARIES DOMAIN (LBD) motifs broadly enriched in co-accessible edges with similar GC-rich palindromic binding sites (FDR < 0.05; Figures 5I and 5J). Analysis of reciprocal TF motifs in maize seedling Hi-C chromatin loops indicated similar enrichment profiles as co-accessible ACRs (SCC = 0.86; Figures S4K and S4L). TCP motifs have been previously implicated in topologically associated domain-like boundaries in *Oryza sativa* and *Marchantia polymorpha,* and the distal edges of chromatin loops in *Z. mays* (Karaaslan et al., 2020; Liu et al., 2017; Peng et al., 2019; Sun et al., 2020). Our results implicate independently evolved TF families with CTCF-like function capable of organizing chromatin architecture through DNA-protein interactions.

Dynamic chromatin accessibility underlies developmental trajectories

The apical domains of maize enclose a pool of undifferentiated meristematic cells that give rise to differentiated cells. To define a *cis*-regulatory catalog of temporal cell fate progressions, we ordered nuclei along pseudotime trajectories for 18 developmental continuums. We then identified ACRs, TF loci, and TF motifs with significant variation across each pseudotime trajectory (Figures 6A and S5; Table S5; STAR Methods). To showcase the power of trajectory analysis to characterize a relatively understudied process, we focused on root phloem companion cell (PCC) development (Figure 6B). We identified 8,004 ACRs, 440 TF motifs, 7,955 genes, and 402 TF loci with differential chromatin accessibility across the PCC trajectory (Figure 6C; STAR Methods). Several known meristem and phloem developmental genes, including *AT-RICH INTERACTIVE DOMAIN-CONTAINING 8*, *SUPPRESSOR OF MAX2 1-LIKE3*, and *SUCROSE TRANSPORTER 1*, were identified among the top differentially accessible genes throughout PCC development (Figure 6D) (Baker et al., 2016; Jiang et al., 2010; Wallner et al., 2017).

Studies of root cell fate decisions have focused on the role of the cell cycle in establishing patterns of asymmetric cell division, as quiescent center (QC) and meristematic cells divide slower than cells in transition and elongation zones (Ten Hove and Heidstra, 2008). To investigate the contribution of cell cycling in PCC development, we annotated nuclei using known cell-cycle marker genes (STAR Methods) (Nelms and Walbot, 2019). A consequence of slower DNA replication, the majority of QC and meristem/initial-like nuclei were in S phase, while differentiated PCCs were largely in G1 (Figure 6E). Ordering nuclei by PCC pseudotime revealed sequential progression of cycle stages within each cell type, implicating the cell-cycle context preceding cell fate transitions during PCC development (Figure 6F). Evaluation of accessibility across pseudotime illustrated a global decrease in chromatin accessibility (F-test: $p < 2.2e-16$; Figure 6G). Thus, cell cycling and cell fate transitions in the context of PCC development accompany global decreases in chromatin accessibility, a consequence we posit is associated with acquisition of specialized functions in PCCs relative to meristematic progenitors.

Evolutionary innovation in root development

To explore the degree of regulatory conservation in root development, we profiled chromatin accessibility in 4,655 nuclei from 7-day-old root tissues of *A. thaliana*, integrated single-nuclei chromatin profiles with published *A. thaliana* root single-cell RNA-seq (scRNA-seq) data (n = 12,606) and constructed eight *cis*-regulatory pseudotime trajectories encompassing vascular, dermal, and ground development (Figures 7A–7C, S6, and S7A; Tables S1 and S6).

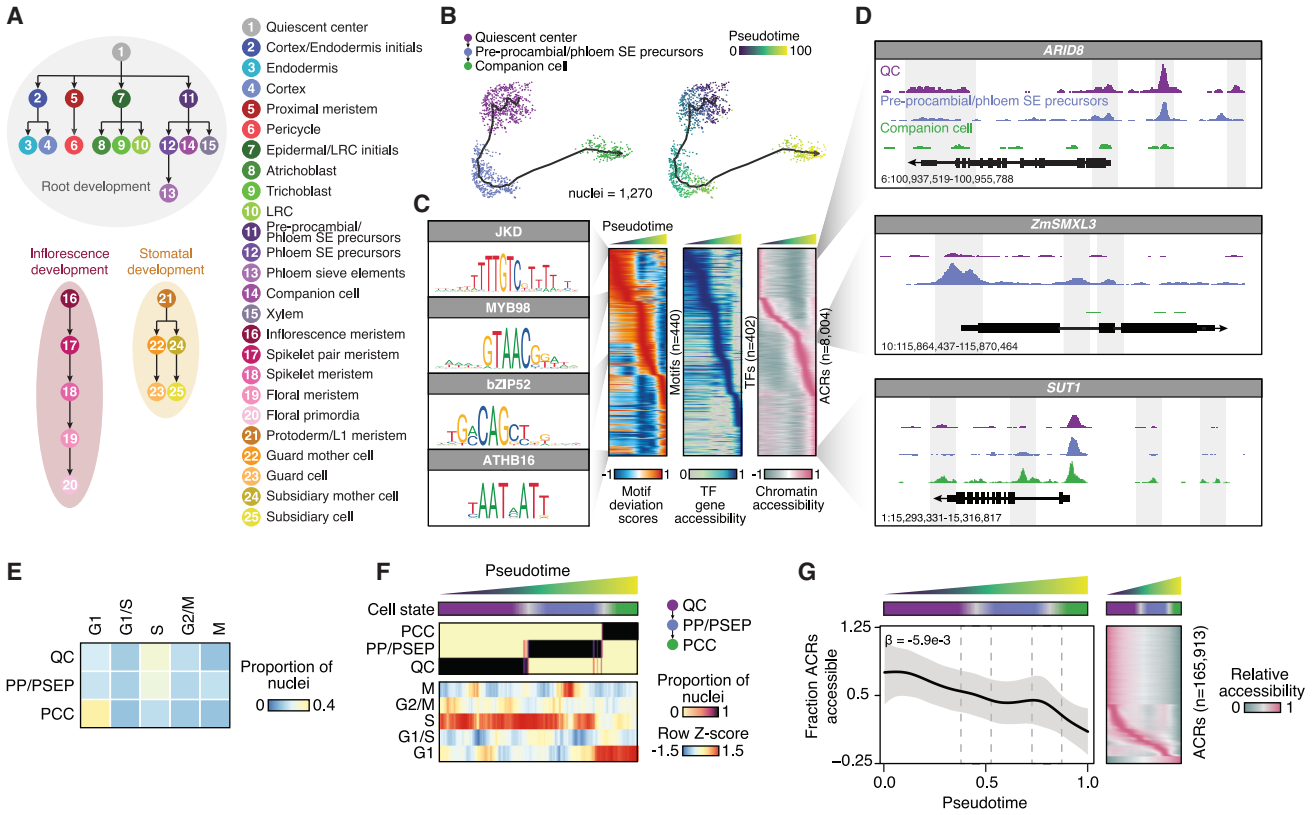


Figure 6. Chromatin accessibility is dynamic across pseudotime

(A) Overview of pseudotime trajectory analysis.

(B) UMAP embedding of the PCC developmental trajectory depicting cell types (left) and pseudotime (right).

(C) Relative motif deviations for 440 TF motifs (left), gene accessibility for 402 TFs (middle), and accessibility of 8,094 ACRs (right) associated with pseudotime (x axis). Four motifs enriched along the trajectory gradient are shown on the left.

(D) Cell-type-specific chromatin accessibility profiles along the developmental trajectory for QC, PP/PSEP, and PCC for three marker genes.

(E) Proportion of cells at various stages of the cell-cycle in QC, PP/PSEP, and PCC annotated clusters.

(F) Top, cell state ordered by pseudotime. Middle, proportion of nuclei with the corresponding cell-type annotation. Bottom, proportion of nuclei with various cell-cycle stage annotations.

(G) Left, average proportion of ACRs accessible across pseudotime. The gray polygon indicates standard deviation. Right, heatmap of relative accessibility (row maximum) for each ACR across pseudotime.

Extending our analysis of PCC development, we first validated the utility of the integrated datasets by visualizing gene expression and accessibility of known marker genes of QC (*WUSCHEL-RELATED HOMEOBOX 5*), procambial (*WUSCHEL-RELATED HOMEOBOX 4*), and PCC (*SUCROSE TRANSPORTER 2*) cell types (Figures 7D and 7E). To enable direct comparison of gene accessibility dynamics in a common space, we aligned *Z. mays* and *A. thaliana* gene orthologs from the PCC trajectories using a dynamic time-warping algorithm (STAR Methods). Consistent with comparative analysis of vascular development in *O. sativa*, *A. thaliana*, and *Solanum lycopersicum* (Kajala et al., 2020), only 206 out of 10,976 putative orthologs were associated (FDR < 0.01) with PCC pseudotime in both species, indicating that the majority of PCC-trajectory-associated genes are unique to each lineage (97% *Z. mays* and 83% *A. thaliana*). However, of the 206 PCC-trajectory-associated orthologs, ~50% (102/206) exhibited similar gene accessibility patterns across pseudotime (Figures 7F, 7G, and S7B). Several orthologs with matching gene accessibility patterns have been previously attributed to PCC development, such as *SCARECROW-LIKE8* (Figure 7H) (Brady et al., 2007). The remaining orthologs (n = 104) clustered into two groups that reflect changes in timing of gene accessibility across the PCC trajectory, underscoring putative functional novelty in PCC development between *Z. mays* and *A. thaliana*.

To reveal innovative *cis* regulation along the pseudotime continuum, we aligned *Z. mays* and *A. thaliana* TF motif profiles associated with PCC progression (Figures S7C and S7D). Of the 440 motifs, 142 demonstrated highly conserved *cis*-regulatory dynamics between species (Figures S7E–S7G). Indeed, the top four motifs ranked by normalized distances (*HOMEOBOX25*, *HOMEOBOX18*, *NAC DOMAIN CONTAINING PROTEIN 55*, and *NAC DOMAIN CONTAINING PROTEIN 83*) have been previously implicated in regulation of hormonal responses and vascular development (STAR Methods) (Jiang et al., 2009; Yamaguchi et al., 2010; You et al., 2019). Gene expression of these TFs from published data was consistent with motif enrichment

Please cite this article in press as: Marand et al., A *cis*-regulatory atlas in maize at single-cell resolution, Cell (2021), https://doi.org/10.1016/j.cell.2021.04.014

CellPress

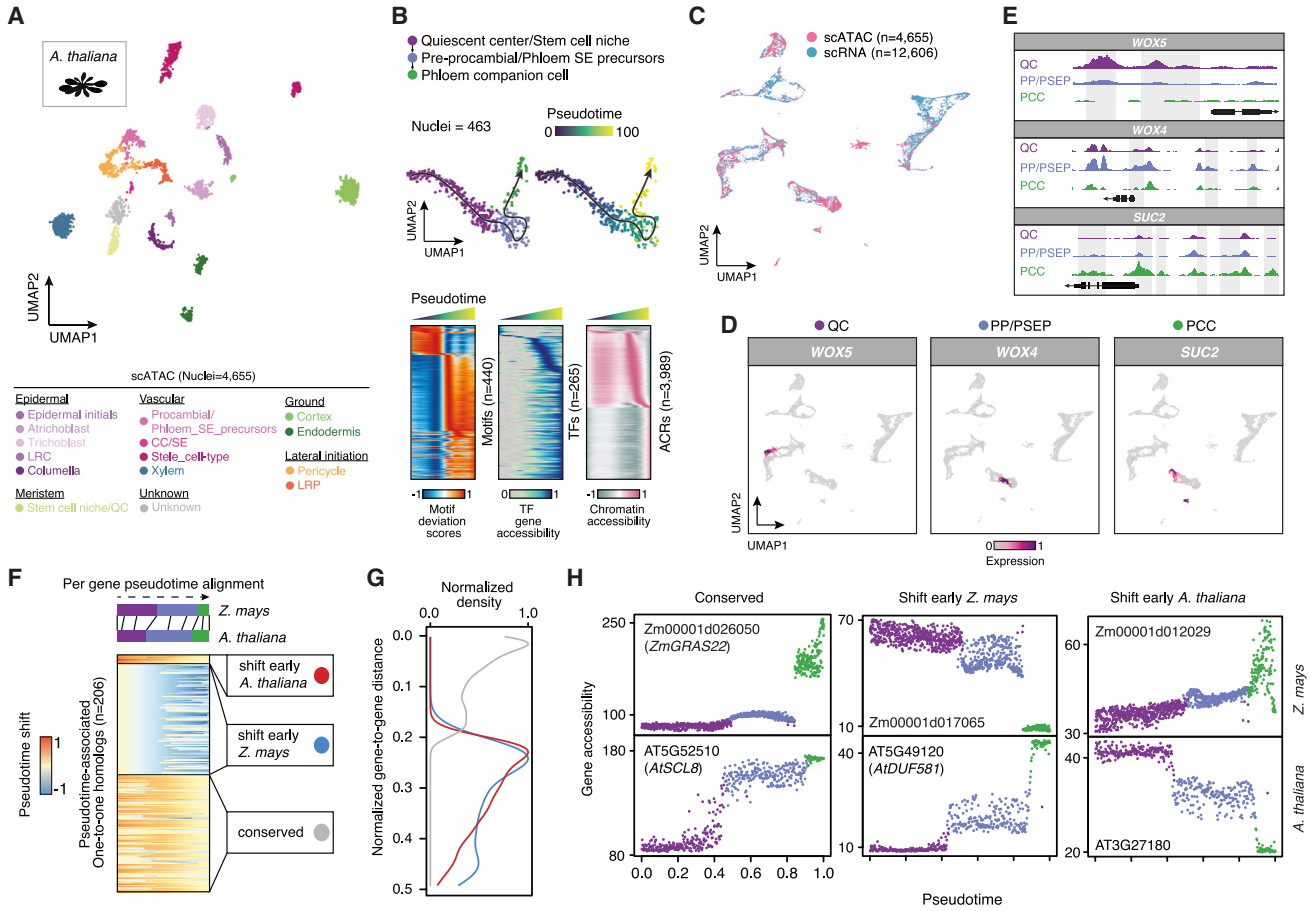


Figure 7. *cis*-regulatory dynamics in *A. thaliana* and *Z. mays* PCC development

(A) UMAP embedding of chromatin accessibility profiles for 4,655 *A. thaliana* root nuclei.

(B) UMAP embedding of *A. thaliana* PCC developmental trajectories depicting cell types (left) and pseudotime progression (right). Motifs, TFs, and ACRs associated with pseudotime are shown as heatmaps.

(C) UMAP embedding of integrated scRNA-seq and scATAC-seq profiles derived from *A. thaliana* roots.

(D) Marker gene expression for QC (*WOX5*), PP/PSEP (*WOX4*), and PCC (*SUC2*).

(E) Cluster-aggregated chromatin accessibility for QC, PP/PSEP, and PCC across three marker genes.

(F) Per-gene pseudotime shift scores from alignments between *Z. mays* and *A. thaliana* PCC development progressions clustered by *k*-means (k = 3).

(G) Distribution of gene-gene distances from *k*-mean groups.

(H) Exemplary one-to-one homologs between *A. thaliana* and *Z. mays* for the three pseudotime shift groups. LRC, lateral root cap; LRP, lateral root primordia. SE, sieve elements.

dynamics in *Z. mays* and *A. thaliana* ontogenies (Figure S7H) (Brady et al., 2007). These findings signify a high degree of conservation in the *cis*-regulatory specification of PCC development between *Z. mays* and *A. thaliana* and provide a framework to realize the core TFs necessary for development of evolutionary conserved plant cells, tissues, and organs.

DISCUSSION

A comprehensive understanding of CREs and TFs across distinct cell types, developmental trajectories, and species is essential for detangling the determinants of cellular identity and the sources of regulatory variation. From an applied perspective, a catalog of CREs at single-cell resolution paves the way for new applications in biotechnology and empowers the generation of *de novo* phenotypic variation in crops. Here, we coupled droplet-based profiling of native nuclei with quasibinomial logistic regression for scATAC-seq data analysis in our species-agnostic software, *Socrates*, to enable fine-scale investigation of *cis*-regulatory variation across diverse plant cell types.

The identification of CREs and TFs with distinct signatures throughout differentiation and within specific cell types provides an inclusive roadmap for interrogating regulatory dynamics across space and time in a global crop. Our results suggest that cellular identity in plants is established by combinatorial TF activity and accessibility of their target binding sites. These findings have major implications for reprogramming of plant cells into desired cell types and/or organs, a long-sought goal. By comparing TF expression and

cognate motif accessibility, we found a number of non-autonomous mobile TFs, an approach that can be extended to other species.

Many of the findings presented here will be particularly useful for innovations in synthetic biology. For example, sequences associated with CRE/PRE activity can be used to tailor the cell-type-specificity of transgene expression cassettes or to direct endogenous gene-silencing via PRC-deposited H3K27me3. TF motifs associated with chromatin architecture could be used to prevent crosstalk among discrete regulatory elements within transgenes. Genome editing of plant CREs results in both loss- and gain-of-function alleles and is proving highly valuable for expanding the phenotypic range of traits associated with yield and architecture (Liu et al., 2021; Rodríguez-Leal et al., 2017). Cell-type-specific CREs and co-accessible ACRs provide a high-confidence list of sequences that can be targeted with genome editing to improve maize performance. These efforts will be further enhanced by focusing on CREs that possess phenotype-associated genetic variation.

Our analysis of *cis*-regulatory dynamics through an evolutionary perspective has implications for future experiments in other species. We revealed that LTR retrotransposons are a significant source of *cis*-regulatory variation and played a central role in shaping regulatory circuitries across evolutionary timescales. We demonstrate that alleles underlying floral cell-type-specific ACRs were historical targets of modern agronomic selection, useful for informing breeding strategies and prioritizing candidates for trait engineering. Comparison of PCC differentiation between maize and *A. thaliana* revealed a greater proportion of TF motifs with consistent spatiotemporal patterns relative to gene orthologs, suggesting regulatory dynamics are key for developmental conservation. These findings indicate that an organism's evolutionary history plays an important role in shaping extant *cis*-regulatory networks, the understanding of which is critical for unlocking novel phenotypic variation.

Despite the extensive analyses presented here, there is much more to discover. We provide pseudotime reconstruction for 26 cell differentiation processes. Researchers have access to preprocessed data matrices partitioned by cell type for convenient reanalysis of cell-type-specific ACRs, genes, and TF motifs for any cell type(s) of interest. To facilitate data exploration, we generated accompanying genome browsers of chromatin accessibility for cell types in both maize and *A. thaliana* (http://epigenome.genetics.uga.edu/PlantEpigenome/index.html). Our analyses can be readily reproduced by leveraging the freely available code implemented throughout the study (https://github.com/plantformatics/maize_single_cell_cis_regulatory_atlas). Researchers analyzing scATAC-seq data can benefit from extensive documentation and tutorials accompanying our R package, *Socrates* (https://github.com/plantformatics/Socrates). We anticipate these data and software will lead to myriad discoveries spanning diverse disciplines, enable innovations in biotechnology, and further the basic understanding of cellular specification and plasticity.

Limitations of study

Although we provide evidence supported by RNA *in situ* hybridization and snRNA-seq for the use of chromatin accessibility as a robust proxy of gene expression, the activity of PREs, limitations in per-nucleus sequencing depth, and dropouts can impact the ability to detect significant marker gene enrichment, particularly for cell types with few known markers. Thus, cell-type annotations should be considered preliminary. We anticipate that the cell-type classifications will become refined as single-cell methods become more widely adopted and as more cells are sequenced across various modalities.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Growth conditions
  - Maize seedlings
  - Maize roots
  - Maize inflorescence
  - Maize axillary buds
  - *Arabidopsis* roots
- METHOD DETAILS
  - Single cell ATAC-seq library preparation
  - Single nuclei RNA-seq library preparation
  - *In situ* hybridizations
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - scATAC-seq raw reads processing
  - Comparing scATAC-seq between fresh and frozen samples
  - Cell calling
  - Detection of multiplet droplets
  - *In silico* sorting
  - ACR identification
  - Nuclei clustering
  - Identification of co-accessible ACRs
  - Estimation of gene accessibility scores
  - Cell-type annotation
  - Cell-cycle annotation
  - snRNA-seq data processing
  - Integration of scATAC-seq and snRNA-seq data
  - STARR-seq analysis
  - Analysis of differential chromatin accessibility
  - Estimates of cell-type specificity
  - GO gene set enrichment analysis
  - Motif analysis
  - Analysis of cell type-specific selection signatures
  - Analysis of co-accessible ACRs
  - Co-accessible ACR interactive capacity
  - Pseudotime analysis
  - Analysis of differential accessibility across pseudotime
  - *A. thaliana* scATAC-seq processing
  - Aligning pseudotime trajectories between *A. thaliana* and Z. mays
  - Additional resources

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2021.04.014.

ACKNOWLEDGMENTS

We would like to acknowledge Dr. Zefu Lu, Dr. William Ricci, Tyler Earp, and Julie Nelson from the UGA flow cytometry core; Noah Workman and Julia Portocarrero from the GGBC; and the GACRC for providing valuable assistance. This study was funded with support from the NSF (IOS-1856627 and IOS-1802848) and the UGA Office of Research to R.J.S., the NSF (IOS-1916804) to A.G., and an NSF postdoctoral fellowship in biology (DBI-1905869) to A.P.M. A.G. and R.J.S. are also supported by NSF collaborative awards to support this research (IOS-2026554 and IOS-2026561).

AUTHOR CONTRIBUTIONS

A.P.M. and R.J.S. designed the research. A.P.M. and Z.C. performed the experiments. A.P.M., Z.C., A.G., and R.J.S. analyzed the data. A.P.M. and R.J.S. wrote the manuscript.

DECLARATION OF INTERESTS

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

Received: December 2, 2020
Revised: March 4, 2021
Accepted: April 7, 2021
Published: May 7, 2021

REFERENCES

Alpert, A., Moore, L.S., Dubovik, T., and Shen-Orr, S.S. (2018). Alignment of single-cell trajectories to compare cellular expression dynamics. Nat. Methods *15*, 267–270.

Andersson, R., and Sandelin, A. (2020). Determinants of enhancer and promoter activities of regulatory elements. Nat. Rev. Genet. *21*, 71–87.

Baker, R.F., Leach, K.A., Boyer, N.R., Swyers, M.J., Benitez-Alfonso, Y., Skopelitis, T., Luo, A., Sylvester, A., Jackson, D., and Braun, D.M. (2016). Sucrose Transporter ZmSut1 Expression and Localization Uncover New Insights into Sucrose Phloem Loading. Plant Physiol. *172*, 1876–1898.

Birnbaum, K., Shasha, D.E., Wang, J.Y., Jung, J.W., Lambert, G.M., Galbraith, D.W., and Benfey, P.N. (2003). A gene expression map of the Arabidopsis root. Science *302*, 1956–1960.

Brady, S.M., Orlando, D.A., Lee, J.Y., Wang, J.Y., Koch, J., Dinneny, J.R., Mace, D., Ohler, U., and Benfey, P.N. (2007). A high-resolution root spatiotemporal map reveals dominant expression patterns. Science *318*, 801–806.

Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., and Greenleaf, W.J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature *523*, 486–490.

Cai, Y., Zhang, Y., Loh, Y.P., Tng, J.Q., Lim, M.C., Cao, Z., Raju, A., Lieberman Aiden, E., Li, S., Manikandan, L., et al. (2021). H3K27me3-rich genomic regions can function as silencers to repress gene expression via chromatin interactions. Nat. Commun. *12*, 719.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421.

Chang, Y.M., Liu, W.Y., Shih, A.C., Shen, M.N., Lu, C.H., Lu, M.Y., Yang, H.W., Wang, T.Y., Chen, S.C., Chen, S.M., et al. (2012). Characterizing regulatory and functional differentiation between maize mesophyll and bundle sheath cells by transcriptomic analysis. Plant Physiol. *160*, 165–177.

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890.

Chuck, G.S., Brown, P.J., Meeley, R., and Hake, S. (2014). Maize SBP-box transcription factors unbranched2 and unbranched3 affect yield traits by regulating the rate of lateral primordia initiation. Proc. Natl. Acad. Sci. USA *111*, 18775–18780.

Clark, R.M., Wagler, T.N., Quijada, P., and Doebley, J. (2006). A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture. Nat. Genet. *38*, 594–597.

Corces, M.R., Granja, J.M., Shams, S., Louie, B.H., Seoane, J.A., Zhou, W., Silva, T.C., Groeneveld, C., Wong, C.K., Cho, S.W., et al.; Cancer Genome Atlas Analysis Network (2018). The chromatin accessibility landscape of primary human cancers. Science *362*, eaav1898.

Crisp, P.A., Marand, A.P., Noshay, J.M., Zhou, P., Lu, Z., Schmitz, R.J., and Springer, N.M. (2020). Stable unmethylated DNA demarcates expressed genes and their cis-regulatory space in plant genomes. Proc. Natl. Acad. Sci. USA *117*, 23991–24000.

Cusanovich, D.A., Hill, A.J., Aghamirzaie, D., Daza, R.M., Pliner, H.A., Berletch, J.B., Filippova, G.N., Huang, X., Christiansen, L., DeWitt, W.S., et al. (2018). A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. Cell *174*, 1309–1324.e18.

Deal, R.B., and Henikoff, S. (2011). The INTACT method for cell type-specific gene expression and chromatin profiling in Arabidopsis thaliana. Nat. Protoc. *6*, 56–68.

Dong, P., Tu, X., Chu, P.Y., Lü, P., Zhu, N., Grierson, D., Du, B., Li, P., and Zhong, S. (2017). 3D Chromatin Architecture of Large Plant Genomes Determined by Local A/B Compartments. Mol. Plant *10*, 1497–1509.

Dorrity, M.W., Alexandre, C., Hamm, M., Vigil, A.-L., Fields, S., Queitsch, C., and Cuperus, J. (2020). The regulatory landscape of Arabidopsis thaliana roots at single-cell resolution. bioRxiv, 2020.2007.2017.204792.

Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. *20*, 238.

Fang, R., Preissl, S., Li, Y., Hou, X., Lucero, J., Wang, X., Motamedi, A., Shiau, A.K., Zhou, X., Xie, F., et al. (2020). SnapATAC: A Comprehensive Analysis Package for Single Cell ATAC-seq. bioRxiv, 615179.

Farmer, A., Thibivilliers, S., Ryu, K.H., Schiefelbein, J., and Libault, M. (2020). The impact of chromatin remodeling on gene expression at the single cell level in < em > Arabidopsis thaliana < /em >. bioRxiv, 2020.2007.2027.223156.

Gage, J.L., White, M.R., Edwards, J.W., Kaeppler, S., and de Leon, N. (2018). Selection Signatures Underlying Dramatic Male Inflorescence Transformation During Modern Hybrid Maize Breeding. Genetics *210*, 1125–1138.

Galli, M., Khakhar, A., Lu, Z., Chen, Z., Sen, S., Joshi, T., Nemhauser, J.L., Schmitz, R.J., and Gallavotti, A. (2018). The DNA binding landscape of the maize AUXIN RESPONSE FACTOR family. Nat. Commun. *9*, 4526.

Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. In arXiv e-prints, pp. arXiv:1207.3907.

Gate, R.E., Cheng, C.S., Aiden, A.P., Siba, A., Tabaka, M., Lituiev, D., Machol, I., Gordon, M.G., Subramaniam, M., Shamim, M., et al. (2018). Genetic determinants of co-accessible chromatin regions in activated T cells across humans. Nat. Genet. *50*, 1140–1150.

Gerstein, M.B., Kundaje, A., Hariharan, M., Landt, S.G., Yan, K.K., Cheng, C., Mu, X.J., Khurana, E., Rozowsky, J., Alexander, R., et al. (2012). Architecture of the human regulatory network derived from ENCODE data. Nature *489*, 91–100.

Gómez-Mena, C., de Folter, S., Costa, M.M., Angenent, G.C., and Sablowski, R. (2005). Transcriptional program controlled by the floral homeotic gene AGAMOUS during early organogenesis. Development *132*, 429–438.

Granja, J.M., Corces, M.R., Pierce, S.E., Bagdatli, S.T., Choudhry, H., Chang, H.Y., and Greenleaf, W.J. (2020). ArchR: An integrative and scalable software package for single-cell chromatin accessibility analysis. bioRxiv, 2020.2004.2028.066498.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017–1018.

Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., and Noble, W.S. (2007). Quantifying similarity between motifs. Genome Biol. *8*, R24.

Hafemeister, C., and Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. *20*, 296.

Heger, P., Marin, B., Bartkuhn, M., Schierenberg, E., and Wiehe, T. (2012). The chromatin insulator CTCF and the emergence of metazoan diversity. Proc. Natl. Acad. Sci. USA *109*, 17507–17512.

Hekselman, I., and Yeger-Lotem, E. (2020). Mechanisms of tissue and cell-type specificity in heritable traits and diseases. Nat. Rev. Genet. *21*, 137–150.

Hofmeister, B.T., and Schmitz, R.J. (2018). Enhanced JBrowse plugins for epigenomics data visualization. BMC Bioinformatics *19*, 159.

Jiang, H., Li, H., Bu, Q., and Li, C. (2009). The RHA2a-interacting proteins ANAC019 and ANAC055 may play a dual role in regulating ABA response and jasmonate response. Plant Signal. Behav. *4*, 464–466.

Jiang, K., Zhu, T., Diao, Z., Huang, H., and Feldman, L.J. (2010). The maize root stem cell niche: a partnership between two sister cell populations. Planta *231*, 411–424.

Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M.C., Wang, B., Campbell, M.S., Stein, J.C., Wei, X., Chin, C.S., et al. (2017). Improved maize reference genome with single-molecule technologies. Nature *546*, 524–527.

Kajala, K., Shaar-Moshe, L., Mason, G.A., Gouran, M., Rodriguez-Medina, J., Kawa, D., Pauluzzi, G., Reynoso, M., Canto-Pastor, A., Lau, V., et al. (2020). Innovation, conservation and repurposing of gene function in plant root cell type development. bioRxiv, 2020.2004.2009.017285.

Karaaslan, E.S., Wang, N., Faiß, N., Liang, Y., Montgomery, S.A., Laubinger, S., Berendzen, K.W., Berger, F., Breuninger, H., and Liu, C. (2020). Marchantia TCP transcription factor activity correlates with three-dimensional chromatin structure. Nat. Plants *6*, 1250–1261.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.R., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods *16*, 1289–1296.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. J. Stat. Softw. *28*, 26.

Langdale, J.A., Lane, B., Freeling, M., and Nelson, T. (1989). Cell lineage analysis of maize bundle sheath and mesophyll cells. Dev. Biol. *133*, 128–139.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. In arXiv e-prints, pp. arXiv:1303.3997.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Liu, C., Wang, C., Wang, G., Becker, C., Zaidem, M., and Weigel, D. (2016). Genome-wide analysis of chromatin packing in Arabidopsis thaliana at single-gene resolution. Genome Res. *26*, 1057–1068.

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

Liu, L., Gallagher, J., Arevalo, E.D., Chen, R., Skopelitis, T., Wu, Q., Bartlett, M., and Jackson, D. (2021). Enhancing grain-yield-related traits by CRISPR-Cas9 promoter editing of maize CLE genes. Nat. Plants *7*, 287–294.

Machanick, P., and Bailey, T.L. (2011). MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics *27*, 1696–1697.

Marand, A.P., Zhang, T., Zhu, B., and Jiang, J. (2017). Towards genome-wide prediction and characterization of enhancers in plants. Biochim. Biophys. Acta. Gene Regul. Mech. *1860*, 131–139.

Minnoye, L., Marinov, G.K., Krausgruber, T., Pan, L., Marand, A.P., Secchia, S., Greenleaf, W.J., Furlong, E.E.M., Zhao, K., Schmitz, R.J., et al. (2021). Chromatin accessibility profiling methods. Nat. Rev. Methods Primers *1*, 10.

Miyashima, S., Roszak, P., Sevilem, I., Toyokura, K., Blob, B., Heo, J.O., Mellor, N., Help-Rinta-Rahko, H., Otero, S., Smet, W., et al. (2019). Mobile PEAR transcription factors integrate positional cues to prime cambial growth. Nature *565*, 490–494.

Nag, A., King, S., and Jack, T. (2009). miR319a targeting of TCP4 is critical for petal growth and development in Arabidopsis. Proc. Natl. Acad. Sci. USA *106*, 22534–22539.

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

Ngan, C.Y., Wong, C.H., Tjong, H., Wang, W., Goldfeder, R.L., Choi, C., He, H., Gong, L., Lin, J., Urban, B., et al. (2020). Chromatin interaction analyses elucidate the roles of PRC2-bound silencers in mouse development. Nat. Genet. *52*, 264–272.

Noshay, J.M., Marand, A.P., Anderson, S.N., Zhou, P., Guerra, M.K.M., Lu, Z., O'Connor, C., Crisp, P.A., Hirsch, C.N., Schmitz, R.J., et al. (2020). Cis-regulatory elements within TEs can influence expression of nearby maize genes. bioRxiv, 2020.2005.2020.107169.

O'Malley, R.C., Huang, S.C., Song, L., Lewsey, M.G., Bartlett, A., Nery, J.R., Galli, M., Gallavotti, A., and Ecker, J.R. (2016). Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. Cell *166*, 1598.

Oka, R., Zicola, J., Weber, B., Anderson, S.N., Hodgman, C., Gent, J.I., Wesselink, J.J., Springer, N.M., Hoefsloot, H.C.J., Turck, F., and Stam, M. (2017). Genome-wide mapping of transcriptional enhancer candidates using DNA and chromatin features in maize. Genome Biol. *18*, 137.

Oka, R., Bliek, M., Hoefsloot, H.C.J., and Stam, M. (2020). In plants distal regulatory sequences overlap with unmethylated rather than low-methylated regions, in contrast to mammals. bioRxiv, 2020.2003.2024.005678.

Peng, Y., Xiong, D., Zhao, L., Ouyang, W., Wang, S., Sun, J., Zhang, Q., Guan, P., Xie, L., Li, W., et al. (2019). Chromatin interaction maps reveal genetic regulation for quantitative traits in maize. Nat. Commun. *10*, 2632.

Phillips, J.E., and Corces, V.G. (2009). CTCF: master weaver of the genome. Cell *137*, 1194–1211.

Pliner, H.A., Packer, J.S., McFaline-Figueroa, J.L., Cusanovich, D.A., Daza, R.M., Aghamirzaie, D., Srivatsan, S., Qiu, X., Jackson, D., Minkina, A., et al. (2018). Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. Mol. Cell *71*, 858–871.e8.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Ravasi, T., Suzuki, H., Cannistraci, C.V., Katayama, S., Bajic, V.B., Tan, K., Akalin, A., Schmeier, S., Kanamori-Katayama, M., Bertin, N., et al. (2010). An atlas of combinatorial transcriptional regulation in mouse and man. Cell *140*, 744–752.

Rebeiz, M., and Tsiantis, M. (2017). Enhancer evolution and the origins of morphological novelty. Curr. Opin. Genet. Dev. *45*, 115–123.

Ricci, W.A., Lu, Z., Ji, L., Marand, A.P., Ethridge, C.L., Murphy, N.G., Noshay, J.M., Galli, M., Mejía-Guerra, M.K., Colomé-Tatché, M., et al. (2019). Widespread long-range cis-regulatory elements in the maize genome. Nat. Plants *5*, 1237–1249.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Rodríguez-Leal, D., Lemmon, Z.H., Man, J., Bartlett, M.E., and Lippman, Z.B. (2017). Engineering Quantitative Trait Variation for Crop Improvement by Genome Editing. Cell *171*, 470–480.e8.

Salvi, S., Sponza, G., Morgante, M., Tomes, D., Niu, X., Fengler, K.A., Meeley, R., Ananiev, E.V., Svitashev, S., Bruggemann, E., et al. (2007). Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. Proc. Natl. Acad. Sci. USA *104*, 11376–11381.

Satpathy, A.T., Granja, J.M., Yost, K.E., Qi, Y., Meschi, F., McDermott, G.P., Olsen, B.N., Mumbach, M.R., Pierce, S.E., Corces, M.R., et al. (2019). Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. Nat. Biotechnol. *37*, 925–936.

Savaldi-Goldstein, S., Peto, C., and Chory, J. (2007). The epidermis both drives and restricts plant shoot growth. Nature *446*, 199–202.

Schep, A.N., Wu, B., Buenrostro, J.D., and Greenleaf, W.J. (2017). chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat. Methods *14*, 975–978.

Smit, A.F.A., Hubley, R., and Green, P. (2013-2015). RepeatMasker Open-4.0. http://www.repeatmasker.org.

Stitzer, M.C., Anderson, S.N., Springer, N.M., and Ross-Ibarra, J. (2019). The Genomic Ecosystem of Transposable Elements in Maize. bioRxiv, 559922.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1888–1902.e21.

Sun, Y., Dong, L., Zhang, Y., Lin, D., Xu, W., Ke, C., Han, L., Deng, L., Li, G., Jackson, D., et al. (2020). 3D genome architecture coordinates trans and cis regulation of differentially expressed ear and tassel genes in maize. Genome Biol. *21*, 143.

Tatematsu, K., Nakabayashi, K., Kamiya, Y., and Nambara, E. (2008). Transcription factor AtTCP14 regulates embryonic growth potential during seed germination in Arabidopsis thaliana. Plant J. *53*, 42–52.

Team, R.C. (2013). R: A language and environment for statistical computing (R Foundation for Statistical Computing).

Ten Hove, C.A., and Heidstra, R. (2008). Who begets whom? Plant cell fate determination by asymmetric cell division. Curr. Opin. Plant Biol. *11*, 34–41.

van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A.J., Burdziak, C., Moon, K.R., Chaffer, C.L., Pattabiraman, D., et al. (2018). Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. Cell *174*, 716–729.e27.

Verweij, W., Spelt, C.E., Bliek, M., de Vries, M., Wit, N., Faraco, M., Koes, R., and Quattrocchio, F.M. (2016). Functionally Similar WRKY Proteins Regulate Vacuolar Acidification in Petunia and Hair Development in Arabidopsis. Plant Cell *28*, 786–803.

Villar, D., Berthelot, C., Aldridge, S., Rayner, T.F., Lukk, M., Pignatelli, M., Park, T.J., Deaville, R., Erichsen, J.T., Jasinska, A.J., et al. (2015). Enhancer evolution across 20 mammalian species. Cell *160*, 554–566.

Wallace, J.G., Bradbury, P.J., Zhang, N., Gibon, Y., Stitt, M., and Buckler, E.S. (2014). Association mapping across numerous traits reveals patterns of functional variation in maize. PLoS Genet. *10*, e1004845.

Wallner, E.S., López-Salmerón, V., Belevich, I., Poschet, G., Jung, I., Grünwald, K., Sevilem, I., Jokitalo, E., Hell, R., Helariutta, Y., et al. (2017). Strigolactone- and Karrikin-Independent SMXL Proteins Are Central Regulators of Phloem Formation. Curr. Biol. *27*, 1241–1247.

Wang, B., Lin, Z., Li, X., Zhao, Y., Zhao, B., Wu, G., Ma, X., Wang, H., Xie, Y., Li, Q., et al. (2020). Genome-wide selection and genetic improvement during modern maize breeding. Nat. Genet. *52*, 565–571.

Welch, J.D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., and Macosko, E.Z. (2019). Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. Cell *177*, 1873–1887.e1817.

Witten, D.M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics *10*, 515–534.

Xiao, J., Jin, R., Yu, X., Shen, M., Wagner, J.D., Pai, A., Song, C., Zhuang, M., Klasfeld, S., He, C., et al. (2017). Cis and trans determinants of epigenetic silencing by Polycomb repressive complex 2 in Arabidopsis. Nat. Genet. *49*, 1546–1552.

Yamaguchi, M., Ohtani, M., Mitsuda, N., Kubo, M., Ohme-Takagi, M., Fukuda, H., and Demura, T. (2010). VND-INTERACTING2, a NAC domain transcription factor, negatively regulates xylem vessel formation in Arabidopsis. Plant Cell *22*, 1249–1263.

You, Y., Sawikowska, A., Lee, J.E., Benstein, R.M., Neumann, M., Krajewski, P., and Schmid, M. (2019). Phloem Companion Cell-Specific Transcriptomic and Epigenomic Analyses Identify MRF1, a Regulator of Flowering. Plant Cell *31*, 325–345.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Zhao, H., Zhang, W., Chen, L., Wang, L., Marand, A.P., Wu, Y., and Jiang, J. (2018). Proliferation of Regulatory DNA Elements Derived from Transposable Elements in the Maize Genome. Plant Physiol. *176*, 2789–2803.

Zheng, L., McMullen, M.D., Bauer, E., Schön, C.C., Gierl, A., and Frey, M. (2015). Prolonged expression of the BX1 signature enzyme is associated with a recombination hotspot in the benzoxazinoid gene cluster in Zea mays. J. Exp. Bot. *66*, 3917–3930.

Please cite this article in press as: Marand et al., A *cis*-regulatory atlas in maize at single-cell resolution, Cell (2021), https://doi.org/10.1016/j.cell.2021.04.014

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical commercial assays		
Chromium Next GEM Single Cell ATAC Library and Gel Bead Kit v1.1	10X Genomics	Cat No./ID: 1000176
Chromium Next GEM Single Cell 3′ Kit v3.1	10X Genomics	Cat No./ID: 1000269
Deposited data		
Raw and analyzed data	This paper	GEO: GSE155178
Raw ATAC-seq data from several maize organs	Ricci et al., 2019	GEO: GSE120304
Raw ATAC-seq data from several maize organs	Crisp et al., 2020	GEO: GSE152046
Raw ChIP-seq data from several maize organs	Ricci et al., 2019	GEO: GSE120304
Raw RNA-seq data from several maize organs	Ricci et al., 2019	GEO: GSE120304
Modern maize breeding signatures of selection	Wang et al., 2020	https://static-content.springer.com/esm/art%3A10.1038%2Fs41588-020-0616-3/MediaObjects/41588_2020_616_MOESM3_ESM.xlsx
Maize GWAS SNPs	Wallace et al., 2014	https://datacommons.cyverse.org/browse/iplant/home/shared/panzea/GWASResults/Wallace_etal_2014_PLoSGenet_GWAS_hits-150112.txt
Maize DAP-seq	Galli et al., 2018	GEO: GSE111857
Arabidopsis DAP-seq	O'Malley et al., 2016	GEO: GSE60143
Maize LTR insertion times	Stitzer et al., 2019	https://github.com/mcstitzer/maize_genomic_ecosystem/blob/master/te_age/B73v4_recovered_ages.txt
Experimental models: Organisms		
Maize: B73	Maize COOP	PI: 550473
Maize: Mo17	Maize COOP	PI: 558532
Arabidopsis: Col-0	ABRC	Stock: CS70000
Software and algorithms		
Cellranger-atac v1.2	CellRanger ATAC (10X Genomics)	https://support.10xgenomics.com/single-cell-atac/software/pipelines/latest/what-is-cell-ranger-atac
MACS2	Zhang et al., 2008	https://github.com/macs3-project/MACS
Samtools	Li et al., 2009	http://www.htslib.org/download/
Seurat	Stuart et al., 2019	https://satijalab.org/seurat/
BEDTools	Quinlan and Hall, 2010	https://bedtools.readthedocs.io/en/latest/
Cicero	Pliner et al., 2018	https://cole-trapnell-lab.github.io/cicero-release/docs/
FIMO,Tomtom,MEME	Grant et al., 2011	https://meme-suite.org/meme/
chromVAR	Schep et al., 2017	https://github.com/GreenleafLab/chromVAR
liger	Welch et al., 2019	https://github.com/welch-lab/liger
Socrates	This paper	https://github.com/plantformatics/Socrates
Analysis code	This paper	https://github.com/plantformatics/maize_single_cell_cis_regulatory_atlas

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Robert J. Schmitz (schmitz@uga.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

All raw and processed data associated with this study has been deposited in the NCBI GEO database under accession code GEO: GSE155178. The genome-wide chromatin accessibility cell-type profiles for *Arabidopsis thaliana* and *Zea mays* are available in a JBrowse genome browser (Hofmeister and Schmitz, 2018): http://epigenome.genetics.uga.edu/PlantEpigenome/index.html. R code used throughout the analysis can be found freely available in the following GitHub repository: https://github.com/plantformatics/maize_single_cell_cis_regulatory_atlas. We also released an *R* package for pre-processing, normalization, clustering, and other downstream analytical steps into streamlined toolkit of scATAC-seq data that can be found in the following GitHub repository: https://github.com/plantformatics/Socrates.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Growth conditions

For libraries derived from seedlings, kernels from genotypes B73 and Mo17 were obtained from USDA National Plant Germplasm System (https://npgsweb.ars-grin.gov) and sown in Sungro Horticulture professional growing mix (Sungro Horticulture Canada Ltd.). Soil was saturated with tap water and placed under a 50/50 mixture of 4100K (Sylvania Supersaver Cool White Delux F34CWX/SS, 34W) and 3000K (GE Ecolux w/ starcoat, F40CX30ECO, 40W) lighting. Seedlings were grown under a photoperiod of 16 hours of light, eight hours of dark. The temperature was approximately 25°C during light hours with a relative humidity of approximately 54%.

Maize seedlings

Above ground seedling tissues were harvested between 8 and 9 AM six days (V1-stage) after sowing. We used both fresh (B73/Mo17 pooled) and flash frozen (B73 only) seedling tissue to construct scATAC-seq libraries (Table S1).

Maize roots

Maize root samples were obtained as follows: B73 kernels were sterilized with 70% EtOH treatment for 5 minutes. After removing the ethanol solution, kernels were suspended with 50% bleach for 30 minutes, followed by five washes with autoclaved Milli-Q water. Sterilized kernels were then sown onto mesh plates with half strength MS (Phytotech laboratories, catalog: M519) media and wrapped in Millipore tape. Plates were incubated in a Percival growth chamber with a photoperiod of 16 hours of light, eight hours of dark. The growth chamber temperature was set to 25°C with a relative humidity of approximately 60%. Apical root tips (bottom 2 cm) of seminal and primary root samples were harvested six days (V1-stage) after sowing between 8 and 9 am. Crown root samples (21 days after sowing) were derived from the three developmental zones of greenhouse grown B73 plants between 8 and 9 am and rinsed with sterile water 3 times.

Maize inflorescence

Data generated from young inflorescence (ear and tassel primordia) were derived from B73 maize grown in the greenhouse. Inflorescence primordia were extracted from shoots harvested approximately one month (V7-stage, 2-4 mm) after sowing, between 8 and 9 AM. Inflorescence primordia between three and eight millimeters from the base to the apical tip were placed in sterile water and used for nuclei isolation.

Maize axillary buds

Axillary buds (~30 samples per library) were taken from B73 maize plants grown in the greenhouse at approximately the same developmental stage (V7) as tassel and ear primordia.

Arabidopsis roots

Seven-day old *A. thaliana* roots were prepared similarly as for maize with the exception of deriving nuclei from whole roots.

METHOD DETAILS

Single cell ATAC-seq library preparation

Each library was prepared by mixing at least three independent biological samples (3-4 seedlings, 3 tassel or ear primordia, 12-14 root tips, 12-14 crown root samples, ~30 axillary buds, and 100-200 *A. thaliana* whole roots). One scATAC-seq library (B73 seedling)

was derived from flash frozen tissue (liquid nitrogen, followed by 7-day −80°C storage), while the remaining libraries were constructed with freshly harvested tissue (Table S1).

To isolate individual plant nuclei, fresh or flash frozen tissue from multiple biological samples were placed on Petri dishes and vigorous chopped with a No. 2 razor blade for two minutes in ~500 uL LB01 buffer (15mM Tris pH 7.5, 2mM EDTA, 0.5mM Spermine, 80mM KCl, 20mM NaCl, 15mM 2-ME, 0.15% TrixtonX-100). Homogenized tissue was then filtered through two layers of miracloth, stained with DAPI to a final concentration of ~1uM and loaded onto a Beckman Coulter MoFlo XDP flow cytometer instrument. A total of 120,000 nuclei were sorted for each sample across four catch tubes (30,000 nuclei each) containing 200 uL LB01. Isolated nuclei were spun down in a swinging-bucket (5 minutes, 500 *rcf.*) centrifuge resuspended in 10uL LB01, pooled, and then visualized on a hemocytometer with a fluorescent microscope. Nuclei suspensions were then spun down (5 minutes, 500 *rcf.*) and resuspended in diluted nuclei buffer (10X Genomics) to a final concentration of 3,200 nuclei per uL and used as input for scATAC-seq library preparation (5 uL; 16,000 nuclei total). Samples were kept on ice for all intermittent steps. For B73/Mo17 mixed library, we pooled 8,000 nuclei from both B73 and Mo17 that were independently isolated. Single-cell ATAC-seq libraries were constructed according to the manufacturer's instruction (10X Genomics, catalog: 1000176). Libraries were sequenced with Illumina NovaSeq 6000 in dual-index mode with eight and 16 cycles for i7 and i5 index, respectively.

Single nuclei RNA-seq library preparation

We prepared snRNA-seq libraries from two biological replicates, each composed of three independent 7-day old B73 seedlings. Seedlings were vigorously chopped with a No. 2 razor blade on a Petri dish in 500 uL of nuclei isolation buffer (Phosphate-Buffered Saline [PBS; ThermoFisher], 500U SUPERase RNase inhibitor [Invitrogen], 1mM 1,4-Dithiothreitol [DTT; Millipore Sigma], and 0.05% Triton X-100 [Millipore Sigma]). Homogenized tissue in nuclei isolation buffer was filtered through a 40-um cell strainer (pluriSelect) and spun at 500 *rcf.* for 5 minutes. The supernatant was discarded, followed by two more wash (500 uL nuclei isolation buffer) and centrifugation steps (500 *rcf.* for 5 minutes), discarding the supernatant and resuspending in 10 uL nuclei isolation buffer lacking Triton X-100. The concentration of nuclei in solution was estimated on a hemocytometer under a fluorescent microscope and adjusted to 2,000 nuclei per uL with nuclease-free water. Single-nuclei RNA-seq libraries were prepared from a total of 16,000 nuclei per library following the manufactures instructions for the Single Cell Gene Expression 3′ V3 library kit (10X Genomics, catalog: 1000269). Libraries were sequenced on an Illumina NovaSeq 6000 in dual-index mode.

In situ hybridizations

3-4mm tassel and ear primordia and young seedlings from the maize B73 inbred line were dissected and fixed in a cold paraformaldehyde acetic acid solution (4% PFA) for 48 hours. Following dehydration through a graded ethanol series and clearing of the tissue with a Histo-clear II solution (Electron Microscopy Sciences), samples were embedded using Paraplast Plus tissue embedding media (McCormick Scientific). 8mm sections were hybridized at 56°C with antisense probes labeled with digoxigenin (DIG RNA labeling mix, Roche), and detected using NBT/BCIP (Roche). Probes were synthesized by *in vitro* transcription (T7 RNA polymerase, Promega) of PCR products obtained from embryo cDNA or from digested full-length cDNA clones. The vectors and primers used for probe design are listed in Table S7.

QUANTIFICATION AND STATISTICAL ANALYSIS

scATAC-seq raw reads processing

The following data processing was performed using each tissue and/or replicate independently unless noted otherwise. Raw BCL files were demultiplexed and convert into fastq format using the default settings of the 10X Genomics tool *cellranger-atac mkfastq* (v1.2.0). Partial raw read processing (adaptor/quality trimming, mapping and barcode attachment/correction) was carried out with *cellranger-atac count* (v1.2.0) using AGPv4 of the maize B73 reference genome (Jiao et al., 2017). Properly paired, uniquely mapped reads with mapping quality greater than 10 were retained using *samtools view* (v1.6; -f 3 -q 10) and by filtering reads with XA tags (Li et al., 2009). Duplicate fragments were collapsed on a per-nucleus basis using *picardtools* (http://broadinstitute.github.io/picard) *MarkDuplicates* (v2.16; BARCODE_TAG = CB REMOVE_DUPLICATES = TRUE). Reads mapping to mitochondrial and chloroplast genomes were counted for each barcode, then excluded from downstream analysis. We removed reads representing potential artifacts by excluding alignments coincident with a blacklist of regions composed of low-complexity and homopolymeric sequences (*RepeatMasker* v4.07) (Smit et al., 2013-2015), nuclear sequences with homology (greater than 80% identity and coverage) to mitochondrial and chloroplast genomes (*BLAST+* v2.7.1) (Camacho et al., 2009), regions exhibiting Tn5 integration bias from Tn5-treated genomic DNA (1-kb windows with greater than 2-fold coverage over the genome-wide median), and potential collapsed sequences in the reference (1-kb windows with greater than 2-fold coverage over the genome-wide median using ChIP-seq input). Genomic Tn5 and ChIP input data were acquired from Ricci, Lu and Ji et al. BAM alignments were then converted to single base-pair Tn5 integration sites in BED format by adjusting coordinates of reads mapping to positive and negative strands by +4 and −5, respectively, and retaining only unique Tn5 integration sites for each distinct barcode. Sequencing saturation was calculated as the proportion of unique reads relative to the estimated library complexity output by the *MarkDuplicates* function apart of picardtools.

Comparing scATAC-seq between fresh and frozen samples

Flash-freezing samples prior to nuclei isolation and scATAC-seq library preparation may expand the ability to profile various tissues and developmental stages that are otherwise difficult to sample, time, or coordinate with other experiments. Thus, we compared various single-cell and bulk ATAC-seq quality metrics between libraries prepared from fresh and frozen B73 seedlings. Data quality in fresh versus frozen library preparations were not significantly different (Welch's t test; P value = 0.14) in a comparison between the distributions of unique Tn5 integration sites per nucleus from the two preparations. On a genome-wide scale, ACRs were identified (*macs2*–nomodel–extsize 150–shift −75) from unique Tn5 integration sites reflective of each preparation method and merged (union of ACRs) using *bedtools merge* (Quinlan and Hall, 2010). This merged set of ACRs was used to count the number of unique Tn5 integration sites overlapping ACRs. Raw counts were scaled with the R function, *cpm,* a part of the *edgeR* package (log = T, prior = 5) and quantile normalized using *normalize.quantiles* from the R package "*preprocessCore*." The concordance between fresh and frozen preparation methods was then assessed using Spearman's rho using the *cor* function in base R (Spearman's rho = 0.90).

Cell calling

To identify high-quality nuclei (a term used interchangeably with "barcodes") using the filtered set of alignments, we implemented heuristic cutoffs for genomic context and sequencing depth indicative of high-quality nuclei. Specifically, we fit a smoothed spline to the $\log_{10}$ transformed unique Tn5 integration sites per nucleus (response) against the ordered $\log_{10}$ barcode rank (decreasing per-nucleus unique Tn5 integration site counts) using the *smooth.spline* function (spar = 0.01) from base R (Team, 2013). We then used the fitted values from the smoothed spline model to estimate the first derivative (slope), taking the local minima within the first 16,000 barcodes as a potential knee/inflection point (16,000 was selected to match the maximum number of input nuclei). We set the unique Tn5 library depth threshold to the lesser of 1,000 reads and the knee/inflection point, excluding all barcodes below the threshold. Spurious integration patterns throughout the genome can be representative of incomplete Tn5 integration, fragmented/low-quality nuclei, or poor sequence recovery, among other sources of technical noise. In contrast, high quality nuclei often demonstrate a strong aggregate accessibility signal near TSSs. Therefore, we implemented two approaches for estimating signal-noise ratios in our scATAC-seq data. First, nuclei below two standard deviations from the mean fraction of reads mapping to within 2-kb of TSSs were removed on a per-library basis. Then, we estimated TSS enrichment scores by calculating the average per-bp coverage of 2-kb windows surrounding TSSs, scaling by the average per-bp coverage of the first and last 100-bp in the window (background estimate; average of 1-100-bp and 1901-2000-bp), and smoothing the scaled signal with rolling-means (R package; *Zoo*). Per barcode TSS enrichment scores were taken as the maximum signal within 250-bp of the TSS. Lastly, for each library, we removed any barcode with a proportion of reads mapping to chloroplast and mitochondrial genomes greater than two standard deviations from the mean of the library.

Detection of multiplet droplets

To estimate the empirical proportion of doublets present in our data, we demultiplexed the two-genotype (B73 and Mo17) pooled seedling scATAC-seq sample and assessed the proportion of barcodes reflecting a mixtures of reads derived from both genotypes. Specifically, B73 and Mo17 whole genome short read resequencing data were acquired from PRJNA338953. Paired-end reads were quality and adaptor trimmed with *fastp* (v0.19.5) (Chen et al., 2018) and aligned to the B73 v4 maize reference genome (Jiao et al., 2017) using *BWA mem* (Li, 2013) with non-default settings (-MT 1). Duplicate reads were removed using *samtools rmdup* (Li et al., 2009) (v1.6). The genomic coordinates of short nucleotide variants (SNVs; single nucleotide polymorphisms [SNPs] and small insertions/deletions [INDELs]) for both genotypes were identified using *freebayes* (Garrison and Marth, 2012) (v1.0.0) with non-default settings (–min-repeat-entropy 1–min-alternate-fraction 0.05). Only biallelic SNPs – requiring at least 5 reads per genotype where B73 and Mo17 were homozygous for reference and alternate nucleotides, respectively – were retained. Genotypes were called by modeling allele counts as a binomial distribution with a term accounting for the sequencing error rate, E_t (determined empirically as the fraction of SNPs failing to match either allele), estimating posterior probabilities via Bayes theorem, and assigning the genotype (or mixture of genotypes) with the greatest probability (Equations 1, 2, 3, 4, 5, 6, and 7). Specifically, the probability to observe k out of n SNPs from B73 can be modeled as a binomial distribution for each B73 (A_1), Mo17 (A_2), and doublet barcode state (N) (Equations 1, 2, and 3):

$$P(k|A_1) = \binom{n}{k} \times E_t^{n-k} \times (1 - E_t)^k \quad (1)$$

$$P(k|A_2) = \binom{n}{k} \times (1 - E_t)^{n-k} \times E_t^k \quad (2)$$

$$P(k|N) = \binom{n}{k} \times (0.5)^k \times (0.5)^{n-k} \quad (3)$$

Let $P(A_1|k)$, $P(A_2|k)$, and $P(N|k)$ reflect posterior probabilities for genotypes B73, Mo17, and doublet barcodes given k allele counts from B73; posterior probabilities can be estimated as follows (Equations 4, 5, and 6):

$$P(A_1|k) = \frac{P(k|A_1) \times P(A_1)}{\sum_{i=0}^{n} P(k|A_i) \times P(A_i)} \quad 4$$

$$P(A_2|k) = \frac{P(k|A_2) \times P(A_2)}{\sum_{i=0}^{n} P(k|A_i) \times P(A_i)} \quad 5$$

$$P(N|k) = \frac{P(k|N) \times P(N)}{\sum_{i=0}^{n} P(k|A_i) \times P(A_i)} \quad 6$$

Finally, the genotype called for each barcode was determined as the event with the greatest posterior probability (Equation 7):

$$\max\{P(A_1|k),\ P(A_2|k), P(N|k)\} \quad 7$$

In silico sorting

To provide sufficient sensitivity for peak calling prior to clustering, we followed an *in-silico* sorting strategy to identify crude clusters of similar cells within each organ (Cusanovich et al., 2018). To do so, we generate a binary matrix representing the presence/absence of Tn5 integration sites in 1-kb windows across all cells in a given organ. Bins with less than 1% accessible cells and cells with less than 100 accessible bins were removed. This binary matrix was then transformed using the matrix normalization method term-frequency inverse document-frequency (TF-IDF). Briefly, the TF term was estimated by weighting binary counts at each bin by the total number of bins containing Tn5 integration sites in a given cell, scaling each cell to sum to 100,000, adding a pseudo-count of one, and log transforming the resulting values to reduce the effects of outliers in downstream processing. The IDF term was calculated as the log transformed ratio of the total number of nuclei to the number of nuclei that were marked as accessible for a given bin. We add a pseudo-count of one to the inverse frequency term to avoid taking the log of zero. The TF-IDF scaled matrix was estimated by taking the dot product of the TF and IDF matrices. To enable faster downstream computation, we kept the top 25,000 bins with the greatest TF-IDF variance across nuclei. The reduced TF-IDF matrix was denoised with singular value decomposition (SVD), retaining the 2^{nd} – 11^{th} dimensions (termed Latent Semantic Indexing, LSI). Each row was centered and standardized, capping the values at ± 1.5. Crude clusters were visually identified using ward.D2 hierarchical bi-clustering on the cosine distances of LSI nuclei and bin embeddings.

ACR identification

ACRs were identified by treating each bulk and single-cell ATAC-seq library as a traditional bulk ATAC-seq library. Aligned reads were filtered by mapping quality greater than 10, and duplicate reads were removed via *samtools rmdup.* We then identified ACRs for each library by converting the BAM alignments in BED format, adjusting the coordinates to reflect single-base Tn5 integrations, and running *MACS2* (Zhang et al., 2008) with non-default parameters: --extsize 150 --shift -75 --nomodel --keep-dup all. A final set of ACRs for comparing bulk and aggregate scATAC-seq libraries (Figure S1) was constructed by taking the union of ACRs across all libraries. To leverage the increased sensitivity afforded by cell-type resolved cluster information while ensuring robust reproducibility in ACR identification, we generated pseudo-replicated bulk alignments using the LSI-based crude clusters (see above, "*In-silico* sorting"). Pseudo-replicates were constructed by randomly allocating nuclei from each cluster into two groups, with a third group composed of all cells from the cluster (cluster bulk). These groupings were used to concatenate Tn5 integration sites corresponding to the nuclei from each group into three BED files. ACRs were then identified from the enrichment of Tn5 integration sites from the pseudo-replicate or cluster bulk aggregates using *MACS2* run with non-default parameters: --extsize 150 --shift −75 --nomodel–keep-dup all. ACRs from both pseudo-replicates and the cluster bulk were intersected with BEDtools, retaining ACRs on the conditional intersection of all three groupings (both pseudo-replicates and the cluster bulk) by at least 25% overlap. The remaining ACRs were then redefined as 500-bp windows centered on the ACR coverage summit. To integrate information across all clusters, ACRs from each cluster were concatenated into a single master list. Lastly, overlapping ACRs were filtered recursively to retain the ACR with the greater normalized kernel Tn5 integration density as previously described (Satpathy et al., 2019).

Nuclei clustering

Starting with a binary nucleus x ACR matrix, we first removed ACRs that were accessible in less than 0.5% of all nuclei, and filtered nuclei with less than 50 accessible ACRs. Inspired by recent developments in modeling single-cell RNA-seq data (Hafemeister and Satija, 2019), we developed a regularized quasibinomial logistic framework that overcomes noise inherent to sparse, binary scATAC-seq data by pooling information across ACRs while simultaneously removing variation due to technical effects, particularly those stemming from differences in barcode sequencing depths. First, a subset of 5,000 representative ACRs selected by kernel density sampling of ACR usage (fraction nuclei that are accessible at a given ACR) were used to model the parameters of each ACR, using

Please cite this article in press as: Marand et al., A *cis*-regulatory atlas in maize at single-cell resolution, Cell (2021), https://doi.org/10.1016/j.cell.2021.04.014

ACR usage as a covariate in a generalized linear model. Specifically, the expected accessibility of an ACR, y_i, can be estimated with a generalized linear model containing a binomial error distribution and logit-link function, and an overdispersion term with a quasibinomial probability density function (Equation 8).

$$\mathbb{E}(y_i) \sim \beta_0 + \beta_1 log10(t) \quad 8$$

Where t is a vector of the sums of accessible ACRs across cell j (Equation 9):

$$t = \sum_i y_{ij} \quad 9$$

To prevent over-fitting and ensure robust estimates in light of sampling noise, we learned the global regularized model parameters, including overdispersion, using the representative ACRs by fitting each parameter against the $\log_{10}$ fraction of accessible nuclei via kernel regression, resulting in smoothed parameter estimates across the spectrum of ACR accessibility penetrance present in these data. The learned global regularized model parameters were then used to constrain fitted values across all ACRs for each nucleus with a simple affine transformation. To account for technical variation among nuclei (variation in barcode $\log_{10}$ transformed read-depth, in particular) we calculated Pearson residuals for each ACR, scaling the residuals by the regularized dispersion estimate and centering values via mean subtraction, representing variance-stabilized and read-depth normalized values of accessibility for a nucleus at a given ACR. We note that this method is amenable to calculating residuals that account for additional sources of technical variation, including categorical and numeric covariates, that may obscure biological signal, such as batch effects, proportion of mitochondrial reads, etc.

The dimensionality of the Pearson residual matrix was reduced using singular value decomposition (SVD) implemented by the R package *irlba* (Witten et al., 2009), retaining the first 25 left singular vectors scaled by singular values (hereafter referred to as nuclei embeddings), analogous to principal components (PCs) on an uncentered matrix. Nuclei embeddings were then standardized across components and filtered to remove components correlated with barcode read depth (Spearman's rho > 0.7). We further reduced the dimensionality of the nuclei embedding with Uniform Manifold Approximation Projection (UMAP) via the R implementation of *umap-learn* (min_dist = 0.1, k = 50, metric = "euclidean"). Nuclei were clustered with the *Seurat* v3 (Stuart et al., 2019) framework and Louvain clustering on a k = 50 nearest neighborhood graph at a resolution of 0.02 with 100 iterations and 100 random starts. Clusters with aggregated read depths less than 1.5M were removed. To filter outliers in the UMAP embedding, we estimated the mean distance for each nucleus with its k (k = 50) nearest neighbors and removed nuclei greater than 3 standard deviations from the mean.

We observed fine-scale heterogeneity within major clusters, thus we repeated our clustering pipeline for each major cluster independently by partitioning the SVD embedding into the top 20 components, L2 normalizing nuclei embeddings across components, and projecting the L2-normalized embeddings into the UMAP space. Subclusters of nuclei were identified by Louvain clustering on the L2 normalized SVD embedding (resolution set manually, range = 0.6 – 1.0) with 20 nearest neighbors, filtering outlier nuclei more than 2 standard deviations from the mean distance of 25 nearest neighbors within each cluster.

For analysis of chromatin accessibility across clusters, we assembled a matrix of clusters by ACRs by aggregating the number of single-base resolution Tn5 integration sites from nuclei within the same cluster for each ACR, analogous to normalizing by the proportion of reads in peaks for each cluster. To account for differences in read depth and other technical factors, the raw counts were transformed with *edgeR's* "cpm" (log = T, prior.count = 5) as previously described (Corces et al., 2018). Log-transformed ACR coverage scores were quantile normalized using "normalize.quantiles" with the R package, *preprocessCore.* Finally, to aid data visualization, we estimated per ACR Z-scores across clusters by mean subtraction and standardization (identical to row-wise execution of the R function, "scale").

Identification of co-accessible ACRs

Recent experiments of population-level chromatin accessibility found that pairwise correlations of accessibility among ACRs recapitulates higher-order chromatin interactions observed in Hi-C and other chromatin architecture experiments (Gate et al., 2018). A similar framework was applied to populations of single cells, which showed that co-accessible ACRs are typically more conserved and functionally associated (Buenrostro et al., 2015). To identify potentially functional co-accessible ACRs, we applied a recently developed method, *Cicero* (Pliner et al., 2018), that estimates regularized correlation scores (ranging from −1 to 1) among nearby ACRs with graphical LASSO to penalize potential interactions by physical distances. Using the binary nuclei x ACR matrix as input, we subset nuclei by their subcluster IDs and estimated co-accessibility among ACRs within 500-kb for each of the 92 clusters, independently. *Cicero* was run by applying a background sample of 100 random regions, and 15 nuclei pseudo-aggregates based on k-nearest-neighbors derived from the UMAP coordinates. To control the false discovery rate (FDR) of co-accessible ACR calls, we shuffled the nuclei x ACR matrix such that the total number of reads per ACR and reads per nucleus were identical to the original matrix. We then repeated co-accessible ACR identification with the shuffled matrix, keeping the original parameters to *Cicero* unchanged. Empirical FDR cluster-specific cut-offs were constructed by identifying the minimum positive co-accessibility score in the background where the FDR < 0.05. Co-accessible links below cluster-specific thresholds were removed. Co-accessible ACRs passing thresholds were compared with previously published HiC and HiChIP datasets derived from maize seedling and pistillate inflorescence primordia (Ricci et al., 2019).

Please cite this article in press as: Marand et al., A *cis*-regulatory atlas in maize at single-cell resolution, Cell (2021), https://doi.org/10.1016/j.cell.2021.04.014

Estimation of gene accessibility scores

Chromatin accessibility at TSSs and gene bodies exhibit marked correlation with transcription output in bulk samples (Extended Data Figure 3F). To aid the identification of marker genes underlying distinct cell-types, we used *Cicero* to estimate gene activity scores. *Cicero* models gene activity as a weighted accessibility score that integrates both proximal and distal regulatory elements linked to a single gene by co-accessibility analysis (see above section "Identification of co-accessible ACRs"). Relative gene accessibility scores per nucleus were estimated by taking a weighted average (3:1, gene body score to proximal/distal activity) of the scaled number of reads mapping to gene bodies for each barcode (summing to 1) with the *Cicero* estimate of gene activity derived from ACRs mapping to 1-kb upstream of gene TSSs and their associated distal ACRs linked by co-accessible ACRs passing FDR < 0.05 thresholds (connected ACRs were constrained to a minimum and maximum intervening distance of 1- and 500-kb, respectively). These weighted gene accessibility scores were rescaled such that gene accessibility scores for a given nucleus summed to 1.

Relative gene accessibility scores exhibited a bimodal distribution with relative gene accessibility values near zero resembling low or non-expressed genes. We applied a Gaussian mixture-model (two distributions) based scaling step per cluster to reduce noise introduced by genes with low gene accessibility. Briefly, the average gene accessibility across nuclei was fit to a two distribution Gaussian mixture model in each cluster using the R package *mclust*. We estimated cluster-specific scaling parameters determined as the 5% quantile of non-zero gene accessibility values of genes from the Gaussian distribution with the larger mean, for each cluster. This parameter was then used to scale gene accessibility scores for all genes in each nucleus within the cluster. Scaled gene accessibility scores were rounded to the nearest integer and normalized across all nuclei and clusters using nucleus-specific size factors estimated as the total gene accessibility of a nucleus divided by the exponential of the mean of log-transformed gene accessibility sums across nuclei. To aid visualization, we smoothed normalized gene accessibility scores by estimating a diffusion nearest neighbor graph (k = 15) using the SVD embedding with 3 steps similar to previously proposed methods (Fang et al., 2020; van Dijk et al., 2018). Downstream analyzes based on binarized gene accessibility were conducted by simply converting normalized (non-smoothed) accessibility scores to 1 for all positive values.

Cell-type annotation

To identify and annotate cell types for each barcode, we identified marker genes known to localize to discrete cell types or domains expected in the sampled tissues/organs based on extensive review of the literature (Table S2). To enable gene accessibility comparisons among clusters, we generated three pseudo-replicates for each cluster by resampling nuclei within the cluster such that all cluster pseudo-replicates contained the mean number of nuclei across clusters (number of nuclei per pseudo-replicate = 552) without replacement when possible. To identify genes with increased accessibility relative to other clusters, we constructed a reference panel with three pseudo-replicates by uniformly sampling nuclei without replacement from each organ (number of nuclei per organ = 92), with a total of 552 nuclei per reference panel pseudo-replicate. We then aggregated read counts across nuclei for each gene and pseudo-replicate. Using the *DESeq2* R package, we identified genes with significantly different (FDR < 0.01) accessibility profiles between each cluster and the reference panel.

The list of significantly differentially accessible genes was filtered to retain the genes on our list of cell type specific markers. We initially ranked the top three marker genes in each cluster by their test statistics. To account for clusters containing small proportions of contaminating nuclei of a different cell type, we adjusted the test statistics using a previously described method (Cusanovich et al., 2018), effectively scaling marker activity scores by the proportion of nuclei in the cluster that were derived from an organ in which the marker gene *i* is an expected cell type. Clusters where the top three markers corresponded to the same cell type were annotated with the consensus cell type.

As an independent method for cell-type annotation, we devised a resampling and normalization procedure on the $\log_2$ fold-change values of marker genes to evaluate cell-type enrichment across all possible cell types for each cluster, normalizing enrichment scores by random permutations accounting for different numbers of markers associated with each cell type. Briefly, starting with differential gene accessibility information for each cluster, we iterated over all cell types, extracting markers associated with the cell type of interest. Then, we summed the $\log_2$ fold-changes values of all markers and multiplied the sum by the proportion of markers passing heuristic thresholds (fold-change > 2 and FDR < 0.01). This score was subtracted by the average of 1,000 random permuted scores from combinations of markers from the remaining cell types (selecting the same number of random genes as the cell type of interest) and divided by the standard deviation of the permuted scores. Cell-type enrichment scores in each cluster were scaled from zero to one by dividing each cell-type enrichment score by the maximum scores across possible cell types. This approach is effective in normalizing differences arising from varying numbers of markers specified for each cell type. Additionally, cell-type annotation scores for clusters with mixed or unknown identity are approximately equally distributed, thus controlling ascertainment bias stemming from marker gene selection. Stated differently, an advantage of this approach is that clusters corresponding to cell types with few or no markers in the tested list are left unassigned as their enrichment scores do not deviate significantly from background levels. Finally, scaled cell-type enrichment scores greater than 0.9 were taken as possible annotations and intersected with putative cell-type labels from the marker ranking approach described above.

For clusters with ambiguous marker gene labels, we developed a logistic regression classifier to identify putative cell types based on whole-genome gene accessibility scores of well-annotated cells. First, we counted the number of Tn5 integration sites per cell overlapping 2-kb upstream to 500-bp downstream of each gene. Read counts were transformed by trimmed mean of M-values (TMM) to enable intra and inter-nucleus comparisons using *edgeR* (Robinson et al., 2010), scaling gene accessibility scores in each nucleus with counts per million. Next, we estimated cell-type enrichment scores for each nucleus by calculating the mean accessibility scores of markers for a

Please cite this article in press as: Marand et al., A *cis*-regulatory atlas in maize at single-cell resolution, Cell (2021), https://doi.org/10.1016/j.cell.2021.04.014

given cell type, subtracting the mean background signal defined as 1,000 sets of averaged randomly sampled genes (each set had the same number of genes as the number of markers), divided by the standard deviation of the background signal. Enrichment scores for each nucleus were transformed into a probability distribution by dividing by the sum of cell-type enrichment scores. For each nucleus, we compared the top two most likely cell types, retaining nuclei where the top predicted cell type had a two-fold greater probability than the next most likely assignment. We used these high-confidence cells to train a regularized logistic multinominal classifier with the R package, *glmnet*. Cell-type classifications with less than 10 nuclei in the training set were excluded. We used a LASSO L1 penalty to regularize the logistic classifier, modeling the training set of nuclei as observations and TMM gene accessibility scores as variables. We balanced observations by weighting by the inverse frequency of cell types in the training set. The model was trained with 10-folds and evaluated by testing a 20% hold-out set of nuclei. The predicted cell type for each nucleus in the atlas was taken as the cell type with the greatest probability if the probability ratio between the best and next best assignment was greater than five-fold, otherwise labeled as 'unknown'. Using these per-cell assignments, we defined subclusters as the majority cell type if greater than 50% of nuclei in the cluster were in agreement, labeling clusters with two or more majority cell types as 'mixed' and all other clusters as 'unknown'. All cell-type labels from these three automated approaches were manually reviewed by careful evaluation with UMAP gene accessibility score embeddings and cluster aggregated coverages for all marker genes and refined *ad hoc*.

Cell-cycle annotation

Cell cycle annotation was performed similarly as cell-type annotation. Briefly, we acquired cell-cycle marker genes from Nelms et al., 2019, selecting 35 markers at random for each cell stage (Nelms and Walbot, 2019). The rationale behind selecting equivalent numbers of markers per stage was to prevent biasing cell cycle annotations to cycle stages with more markers, while 35 markers was the minimum gene count across all stages (mitosis). For each stage, we subset the nuclei by gene accessibility (TMM) matrix by the cognate stage, and summed accessibility scores for each nucleus. This cell-cycle stage score was then standardized using the mean and standard deviation of 1,000 permutation of 35 random cell-cycle stage genes, excluding the focal stage. Z-scores corresponding to each cell-cycle stage were converted into probabilities using the R function *pnorm*. Per nucleus posterior cell-cycle probabilities were estimated using Bayes theorem with each cell-cycle stage prior probability set to 0.2 (1/5, for five stages: G1, G1/S, S, G2/M, M). The cell-cycle stage with the maximum probability was selected as the most likely cell stage. Nuclei with multiple cell-cycle annotations with equal maximum probability were considered "ambiguous."

snRNA-seq data processing

Raw fastq files from each snRNA-seq seedling library (across two biological replicates) were processed with *cellranger count* v4.0 to align reads to AGPv4 of the maize B73 reference genome (Jiao et al., 2017). BAM files were filtered to remove multiple mapping reads using a mapping quality filter selecting reads with MQ greater than or equal to 30. The number of nuclear, organeller and transcript-derived unique molecular identifiers (UMIs) reads for each barcode were tabulated from the filtered BAM file. Barcodes with less than 1,000 total UMIs and less than 500 genes with at least one UMI were removed. We then estimated the Z-score distributions for the proportion of mitochondrial, chloroplast, nuclear, and transcript derived UMIs across barcodes. Barcodes above 1 standard deviation (Z-score less than 1) from the mean proportion of UMIs derived from mitochondrial and chloroplast genomes were removed. Likewise, barcodes below 1 standard deviation from the mean proportion of UMIs derived from the nuclear genome were removed.

Integration of scATAC-seq and snRNA-seq data

To integrate scATAC-seq and snRNA-seq data into a shared embedding, we input gene accessibility scores and gene expression values from all seedling-derived nuclei passing quality filters described above using *liger* with the function *createLiger* (Welch et al., 2019). Each dataset was normalized, subset by highly variable genes, and scaled using the functions *normalize, selectGenes,* and *scaleNotCenter,* sequentially with default arguments. An integrated non-negative matrix factorization (iNMF) embedding was constructed from the gene by nuclei scATAC-seq and snRNA-seq matrices using *optimizeALS* with default settings (k = 20, lambda = 5). The iNMF embedding was quantile normalized with *quantile_norm* and non-default settings (do.center = FALSE). Louvain clusters from the normalized iNMF embedding were identified at a resolution of 0.25 with *louvainCluster.* To visualize the integrated assays, we used *runUMAP* with non-default settings (n_neighbors = 20, min_dist = 0.01). Differentially accessible and expressed genes per cluster were identified using *runWilcoxon* requiring FDR less than 0.05 and a $\log_2$ fold change greater 0.25 using the integrated embedding (both gene accessibility and expression across all co-embedded nuclei), gene accessibility in isolation (scATAC-seq nuclei only), and gene expression in isolation (snRNA-seq nuclei only). Differentially accessible ACRs from the normalized (with *liger* function *normalize*) sparse ACR by nuclei matrix were identified using identical heuristic thresholds as for gene expression and accessibility.

To impute ACR accessibility in snRNA-seq derived-nuclei and gene expression values in scATAC-seq nuclei, we ran *imputeKNN* from the *liger* package using either the scATAC-seq or snRNA-seq nuclei as reference cells. We then used the imputed gene expression and ACR accessibility matrices, constrained to only differentially accessible ACRs (n = 55,939), to identify significantly associated gene-to-peak linkages with the *liger* function *linkGenesAndPeaks* with non-default settings (dist = 'spearman', alpha = 0.05). To remove potential false positives, we shuffled the imputed ACR and gene nuclei matrices and repeated gene-to-peak linkage identification using the same arguments. We then estimated FDR empirically over a grid of a 100 possible correlation values in both the negative and positive directions by identifying correlation cut-offs that removed 95% of gene-to-peak linkages from the shuffled matrices. We then filtered the non-shuffled gene-to-peak linkages according to the thresholds identified from the empirical FDR estimates.

Please cite this article in press as: Marand et al., A *cis*-regulatory atlas in maize at single-cell resolution, Cell (2021), https://doi.org/10.1016/j.cell.2021.04.014

STARR-seq analysis

Single bp-resolution enhancer activities were available from a previous study (Ricci et al., 2019). Enhancer activity (defined as the $\log_2$ ratio between RNA and DNA input fragments scaled per million) for each ACR was taken as the maximum over the entire ACR. A control set of regions was generated to match each ACR with the following criteria: (*i*) GC content within 5%, (*ii*) physically constrained to within 50-kb of an ACR, and (*iii*) the same length (500-bp) distribution. The same set of control regions was used throughout the analysis.

Analysis of differential chromatin accessibility

Next, we implemented a logistic regression framework based on binarized ACR accessibility scores for assessing the importance of each ACR to cluster membership by estimating the likelihood ratio between logistic models with, and without a term for cluster membership similar to previous approaches (Cusanovich et al., 2018). Specifically, for each cluster, we compared binarized ACR accessibility scores to a reference panel of uniformly sampled nuclei from each organ where the total number of reference nuclei was set to the average number of nuclei per cluster. We then fit two generalized linear logistic regression models (Equations 10 and 11), with and without a term for membership to the cluster of interest.

$$logit(p_{ij}) = u_i + \alpha_j + \beta_j + \varepsilon_i \quad (10)$$

$$logit(p_{ij}) = u_i + \beta_j + \varepsilon_i \quad (11)$$

Where p_{ij} is the probability that ACR *i* is accessible in nucleus *j*, u_i is the proportion of nuclei where ACR *i* is accessible, α_j is the cluster membership of nucleus *j*, β_j is the $\log_{10}$ number of accessible ACRs in nucleus *j* and ε_i is the error term for the i^{th} ACR. We then used a likelihood ratio test to compare the fits of the two models and estimated the false discovery rate (FDR) using the Benjamini-Hochberg method to identify ACRs that were significantly differentially accessible across clusters by conditioning on FDR < 5% and fold-change threshold greater than two. ACRs meeting these criteria with positive Z-scores in nine or fewer clusters (< 10% of clusters) were considered as cluster-specific. Analysis of differential gene accessibility was performed as described in the section titled "Cell-type annotation."

Estimates of cell-type specificity

To estimate specificity scores for all genes, a binary matrix was constructed indicating the genes (rows) passing heuristic thresholds (FDR < 0.05 & log2-fold change > 2) in each cluster (columns) from differential accessibility testing. Then, starting from a matrix of counts per million (CPM) values across clusters and genes, the relative CPM values for a given gene were converted into a probability distribution by scaling by the sum across clusters. Gene probabilities were split according to the heuristic thresholds, (1) clusters where the gene was significantly enriched, and (2) clusters lacking statistical support for enrichment ("non-enriched"), using the above binary matrix. We summed the gene probabilities for enriched clusters and penalized the sum by 1 – proportion of clusters passing heuristic thresholds, resulting in higher weights for genes enriched in fewer clusters. In parallel, we summed the gene probabilities for non-enriched clusters and penalized the sum by 1 – proportion of clusters failing heuristic thresholds. The specificity score (ranging from −1 to 1) for each gene was estimated as the difference between enriched and non-enriched penalized probability sums. Specificity scores were estimated similarly for ACRs. To determine if cell type-specific *a priori* genes were enriched for greater cell type specificity, we estimated the average specificity across all marker genes and compared the average to a null distribution of 10,000 permutations of randomly selected genes (n = 221).

GO gene set enrichment analysis

Gene set enrichment using GO biological process terms was performed using the R package *fgsea.* For each cluster, test statistics were multiplied by the sign of the $\log_2$ fold-change value versus the reference panel. GO terms with gene sets less than 10 and greater than 600 were excluded from the analysis. GO terms were considered significantly enriched at FDR < 0.05 following 10,000 permutations.

Motif analysis

Motif occurrences were identified genome-wide with *fimo* from the MEME suite toolset (Grant et al., 2011) using position weight matrices (PWM) based on DAP-seq data in *A. thaliana* and *Zea mays* (Galli et al., 2018; O'Malley et al., 2016). To identity TF motifs associated with cell type-specific ACRs, we ranked the top 2,000 ACRs in each cell type by Z-scores derived from CPM normalized accessibility values (see section above "Nuclei clustering"). As a reference for comparison, we identified 2,000 "constitutive" ACRs that varied the least and were broadly accessible across clusters. The number of ACRs containing a specific motif was compared to the frequency of constitutive ACRs harboring the same motif using a binomial test for each cell type and motif. To control for multiple testing, we used the Benjamini-Hochberg method to estimate the FDR, considering tests with FDR < 0.05 as significantly different between the focal cell type and constitutively accessible regions. Maize homologs of *A. thaliana* TFs were identified using protein fasta alignments from BLAST+ v2.10.0 with an E-value cut-off of 1e-5. Only fasta sequences classified as transcription factors

from either species were considered during alignment. To narrow the list of putative orthologs based on functional similarity to *A. thaliana* TFs, we filtered alignments with less than 30% identity, removed maize TFs classified as belonging to a different family, and selected the homolog with the greatest Pearson correlation coefficient with respect to the motif deviation score. Motif deviation scores of specific TF motifs among nuclei were estimated using *chromVAR* (Schep et al., 2017) with the non-redundant core plant PWM database from JASPAR2018. The input matrix for *chromVar* was filtered to retain a minimum of 50 accessible nuclei per ACR and barcodes with at least 50 accessible ACRs. We visualized differences in global motif usage per nucleus by projecting deviation scores onto the UMAP embeddings. To determine if patterns of TF motif accessibility from individual nuclei could be used to predict cell-type annotations, we constructed a neural network for multinomial classification using the R package, *caret* (Kuhn, 2008) (method = "multinom") using 80% of nuclei to train, 10-fold cross-validation, averaging error terms across 10 iterations. The nuclei in the 20% withheld group were used to test the model. Sensitivity, specificity and accuracy of the model was evaluated using the function confusionMatrix from *caret*.

To identity *de novo* motifs enriched in accessible but non-transcribed genes, we selected ACRs (n = 15,576) within 1-kb of genes that were accessible (ATAC log2 TPM > 1.5) and non-expressed (mRNA log2 TPM < 1) in at least 10 clusters. We then constructed a set of control regions by randomly sampling ACRs within 1-kb of genes expressed (mRNA log2 TPM > 1) and accessible (ATAC log2 TPM > 1.5) in at least 10 clusters (n = 15,576). Thresholds were set according to empirical distributions for root and ear tissue. *De novo* motif identification was conducted using the discriminative motif discovery workflow of MEME-ChIP (v5.1.1) with default settings using ACRs 1-kb upstream of TSS for (1) accessible and silenced and (2) accessible and expressed genes as positive and negative sequences, respectively (Machanick and Bailey, 2011). Comparison of *de novo* motifs with experimentally identified motifs was performed using TOMTOM from the MEME-suite toolkit (Gupta et al., 2007).

Analysis of cell type-specific selection signatures

Analysis of within-species genetic variation and phenotype-associated GWAS hits was performed as previously described (Ricci et al., 2019). Multi-locus allele-frequency differentiation signals between chronologically sampled elite maize inbred lines were mapped onto ACRs (Wang et al., 2020), where the selection score for an ACR was taken as the maximum XP-CLR value within the 500-bp ACR interval. To identify cell types associated with increased signatures of selection, the top 2,000 ACRs defined by standardized quantile-scaled CPM chromatin accessibility (Z-scores, see above "Nuclei clustering") were identified for each cell type. The mean XP-CLR scores per-cell type were standardized by the mean and standard deviation of randomly sampled ACRs (n = 2,000) without replacement across 1,000 permutations, where each permutation estimates the mean XP-CLR scores of a random subset of 2,000 ACRs from the total list of 165,913 possible ACRs. Enrichment Z-scores were converted into *P*-values using the R function *pnorm* (log.p = T, lower.tail = F) and used to estimate FDR via the Benjamini-Hochberg method with the R function *p.adjust* (method = "fdr").

Analysis of co-accessible ACRs

To enable comparison with previously identified Hi-C and HiChIP loops (Ricci et al., 2019), we constrained the distance between co-accessible ACRs to the same range as loops identified in leaf Hi-C and HiChIP (minimum loop distance = 20-kb). Co-accessible ACRs and Hi-C/HiChIP loops were considered overlapping if both anchors overlapped by at least 50-bp. We compared motif composition of co-accessible ACRs by scoring motif occurrence as binary for each ACR and estimating a Jaccard similarity score on the union of motif sets. Motif similarity scores for co-accessible ACRs in each cell type were compared to a null distribution by repeating Jaccard similarity calculations for non-co-accessible ACR-ACR connections (constraining the null connections to blocks of 1,000 ACR on the same chromosome with the same ACR-ACR distance distribution as co-accessible ACRs) across 1,000 permutations. To identify motifs enriched at co-accessible ACR anchors, we first estimated the proportion of co-accessible ACRs with an identical motif at both anchors for each motif and cell type. Then, we constructed the same number of random ACR-ACR connections as co-accessible ACRs, again estimating the proportion of links with an identical motif at both anchors, building a null distribution over 1,000 random permutations. The estimated proportion of co-accessible ACRs with identical motifs at both anchors for each motif was transformed to a Z-score by subtracting and scaling by the mean and standard deviation of the null distribution. Z-scores were converted to *P*-values using the R function, *qnorm* with non-default parameters (log.p = T, lower.tail = F). FDR values were estimated using *p.adjust* (method = "fdr"). Co-accessible motif scores were plotted as heatmaps using *heatmap.2* by subtracting and dividing observed with expected proportions. Rows and columns were clustered with *hclust* (method = "ward.D2").

Co-accessible ACR interactive capacity

To derive a per-ACR estimate of interactive capacity, we first counted the number of times an ACR was involved in a co-accessible link for each cell type. The average interactive capacity was taken by averaging across all cell types. Significance tests were performed by subsetting ACRs into two groups (with and without enhancer activity, and with and without overlap with GWAS SNPs) and comparing the distributions with Monte Carlo simulation (permutations = 10,000).

Pseudotime analysis

Pseudotime trajectories were constructed similar to previous methods (Granja et al., 2020). Briefly, nuclei were ordered based on the principal component space by fitting a continuous trajectory via a smooth spline on the Euclidean distances of each nuclei to a predefined order of cell types. For feature analysis (ACRs, motifs, and TF activity) across pseudotime, nuclei were sorted by ascending

pseudotime. The ACR x nucleus matrix was filtered to retain differentially accessible ACRs (see section "Analysis of differential accessibility across pseudotime" below) with at least one nucleus defined as accessible. For each ACR, we fit a generalized additive model with the binary accessibility scores as the response and a smoothed pseudotime component as the dependent variable [s(pseudotime, bs = "cs")] with a binomial error term and a logit-link function with *gam* from the *mgcv* R package. Predicted accessibility scores across pseudotime were generated from 500 equally spaced interpolated points covering the range of pseudotime values. Finally, predicted accessibility scores were mean-centered, standardized and constrained to the range ± 1 for each ACR. Model specification for motif deviations and TF gene accessibility analysis was similar to ACR pseudotime analysis with the exception of a Gaussian error distribution, and TF gene accessibility was normalized by the row maximum rather than rescaling on a ± 1 distribution.

Analysis of differential accessibility across pseudotime

To identity differentially accessible ACRs across pseudotime, we fit normalized accessibility residuals (Pearson residuals from a generalized linear logistic regression model with $\log_{10}$ number of accessible ACRs per barcode as the dependent variable, see section "Nuclei Clustering" above) as the response and pseudotime as the dependent variable using a natural spline with six degrees of freedom [ns(pseudotime, df = 6)] from the R package *splines* for each trajectory. We took an F-test based approach for hypothesis testing of differential accessibility across pseudotime by comparing the variance explained by the splined linear model with that of the residuals normalized by degrees of freedom. *P*-values from the model were used to estimate Benjamini-Hochberg FDR values with the R function *p.adjust* (method = "fdr"), where a FDR threshold < 0.05 denoted statistical significance for differentially accessible ACRs across pseudotime. To identify genes and TF motifs with differential accessibility across pseudotime, we fit the linear splined regression model with the normalized gene accessibility scores and motif deviations from each nucleus, respectively, similar to the analysis of ACRs.

A. thaliana scATAC-seq processing

scATAC-seq data derived from *A. thaliana* root nuclei were processed similarly to the scATAC-seq data derived from maize nuclei. Specifically, we processed raw fastq files using *cellranger-atac*, filtered multi-mapped reads (MQ less than 10 and the presence XA:Z: tags), removed PCR duplicates by barcode, filtered barcodes by proportion of Tn5 integration sites mapping to organellar genomes above 1 standard deviation from the mean, and removed barcodes with less than 1000 unique Tn5 integration sites. We used *in silico* sorting to group nuclei by similarity, identify ACRs, estimate residuals with regularized quasibinomial regression from the binary ACR by nuclei matrix, and reduced dimensions with SVD (singular values = 50) similarly as for maize nuclei. We coded library sequence depth per nucleus as a covariate using the *dplyr* function *ntile* with n = 3 and removed additional technical variance with *Harmony* using the SVD matrix as input with non-default settings for a weak correction (tau = 3, nclust = 15, max.iter.harmony = 30, theta = 0, lambda = 10) (Korsunsky et al., 2019). Nuclei were clustered with Louvain clustering (resolution = 1) in the *Harmony* corrected embedding, and projected into an additionally reduced space with UMAP (n_neighbors = 15, min_dist = 0.1).

Aligning pseudotime trajectories between *A. thaliana* and Z. mays

To enable comparison of companion cell development between *A. thaliana* and *Z. mays*, we first identified putative one-to-one orthologs using *OrthoFinder* (v2) (Emms and Kelly, 2019). Gene accessibility scores for 10,976 putative orthologs were imputed using a diffusion-based approach (Fang et al., 2020; van Dijk et al., 2018) and scaled from 0 to 1 across pseudotime for barcodes associated with companion cell development in *A. thaliana* and *Z. mays*. To account for different distributions, pseudotime coverage, and number of barcodes between species, we used the R package, *cellAlign,* that interpolates, scales, and weights gene accessibility scores on a fixed set of (n = 200) equally spaced points (width parameter: winSz = 0.1) from two trajectories to remove technical biases inherent to each dataset (Alpert et al., 2018). For each putative ortholog, we performed global alignment of gene accessibility scores across *A. thaliana* and *Z. mays* pseudotime using the dynamic time warping algorithm with default settings in *cellAlign.* We then extracted the pseudotime shifts, representing the extent of gene accessibility deviation at any given point along the trajectory, for each putative ortholog. We clustered genes into two groups based on pseudotime shifts across companion cell development using k-means clustering. To identify conserved gene accessibility patterns across pseudotime, we clustered the normalized distances between *A. thaliana* and *Z. mays* putative orthologs using a mixture model (G = 2) with the R package, *mclust.* The mixture model identified a bimodal distribution of normalized distances with 0.15 as a natural cut-off for defining conserved accessibility patterns. Putative orthologs with normalized distances less than the cut-off were placed in a third group defined as conserved. The above analysis was repeated with TF motif deviations scores for 440 TF motifs, without the need for ortholog searching as the same TF position weight matrices were used for both species, affording identical TF motif labels.

Additional resources

Cell-type resolved data can be viewed through our public Plant Epigenome JBrowse Genome Browser (Hofmeister and Schmitz, 2018) (http://epigenome.genetics.uga.edu/PlantEpigenome/index.html) by selecting either the *Z. mays* or *A. thaliana* Genome Browser links, followed by the scATAC_celltypes tab in the tracks panel.

CellPress

Supplemental figures

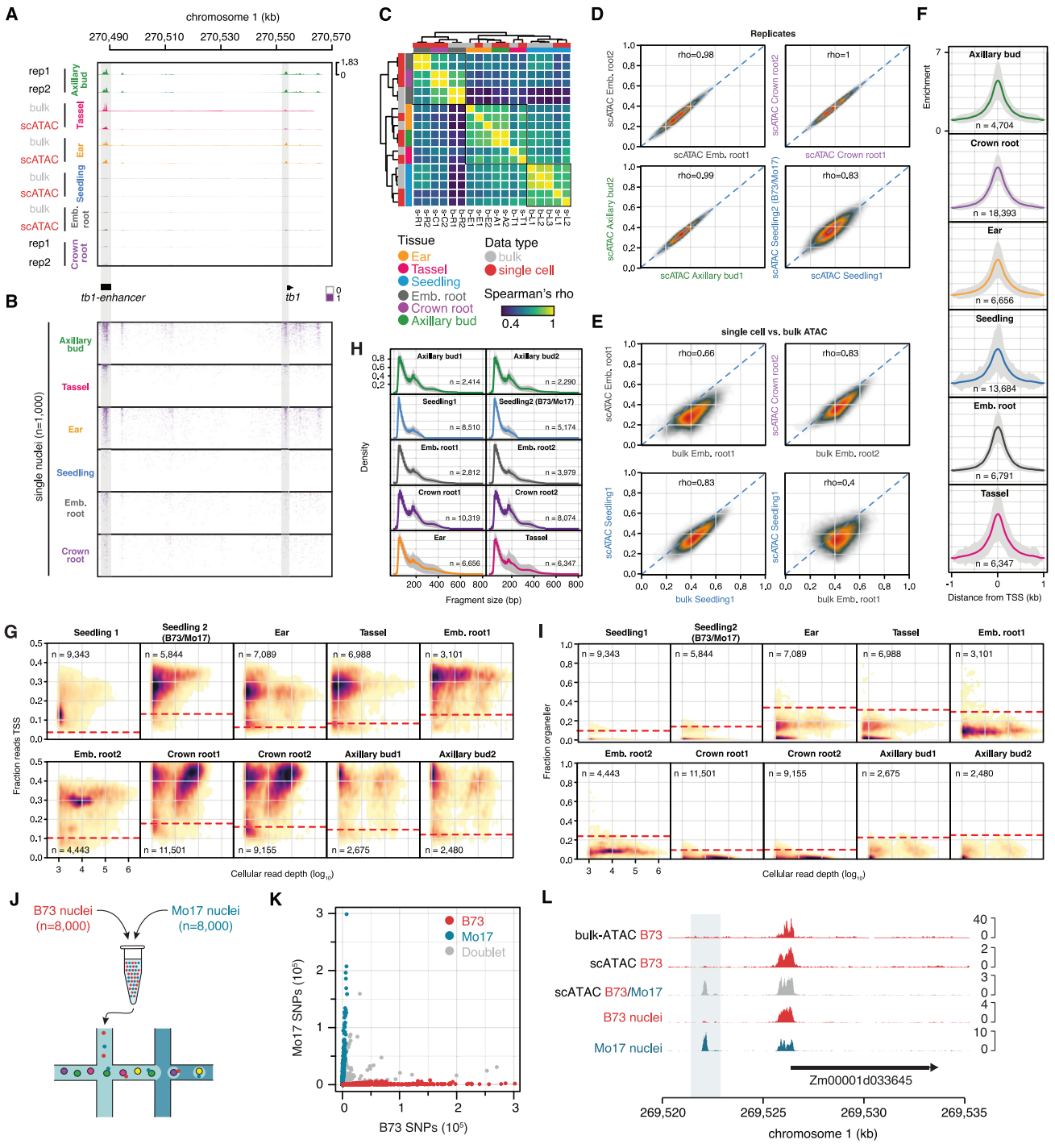


Figure S1. Evaluation and quality control of maize scATAC-seq, related to Figure 1

(A) Genome browser screenshot of chromatin accessibility from bulk and aggregated single-cell ATAC-seq experiments. Chromatin accessibility profiles depict the *tb1* locus and the *tb1* enhancer located approximately 67kb upstream.

(B) Binary accessibility scores from a random selection of 1,000 individual nuclei from each organ.

(legend continued on next page)

(C) Spearman's rho matrix comparing bulk ATAC-seq and aggregate scATAC-seq samples across various organs. Sample codes are shorthand for assay type, sample, and replicate. For example, s-R1 denotes single cell assay for seminal root replicate 1. The term b-L2 denotes a bulk-ATAC assay for seedling replicate 2. Codes are as follows: b, bulk; s, single cell; R, seminal root; C, crown root; E, ear; T, tassel; A, axillary bud; L, seedling. Numbers represent replicate.
(D and E) Comparison of normalized (0-1) read depths at the union of all peaks across bulk and single-cell samples (n = 265,992) between (D) replicated libraries and between (E) bulk and single-cell ATAC-seq assays.
(F) Enrichment plots centered on 2-kb windows surrounding TSSs for barcodes in each tissue. Grey polygons indicate the standard deviation across cells within the noted tissue.
(G) Density scatterplots of $\log_{10}$ transformed barcode read depths (x axis) by the fraction of Tn5 integration sites mapping to within 2-kb of transcription start sites (TSSs). Dashed red lines indicate the threshold of two standard deviations from the mean used to filter lower quality barcodes.
(H) Fragment length distributions for each library. Solid lines indicate the average distribution across cells within the sample. Grey polygons represent the standard deviation across cells in the library.
(I) Density scatterplots of $\log_{10}$ transformed barcode read depths (x axis) by the fraction of Tn5 integration sites derived from organellar sequences (chloroplast and mitochondrial) relative to the total number of unique Tn5 integration sites associated with cognate barcodes. Dashed red lines indicate the threshold of two standard deviations from the mean used to filter lower quality barcodes.
(J) Genotype-mixing experimental schematic.
(K) Scatterplot of per cell B73 and Mo17 SNP counts from a mixed-genotype experiment (V1 seedlings) colored by genotype classification.
(L) Genome browser screenshot of traditional bulk ATAC-seq from 7-day old seedling (row 1), single-cell ATAC-seq from B73 seven-day old seedlings (row 2), pooled B73 and Mo17 nuclei (library ID: Seedling 2) single cell ATAC-seq from seven-day old seedlings (row 3), and the genotype-sorted B73 (row 4) and Mo17 (row 5) alignments after sorting barcodes by genotype calls from the B73-Mo17 scATAC-seq seven-day old seedling sample (row 3).

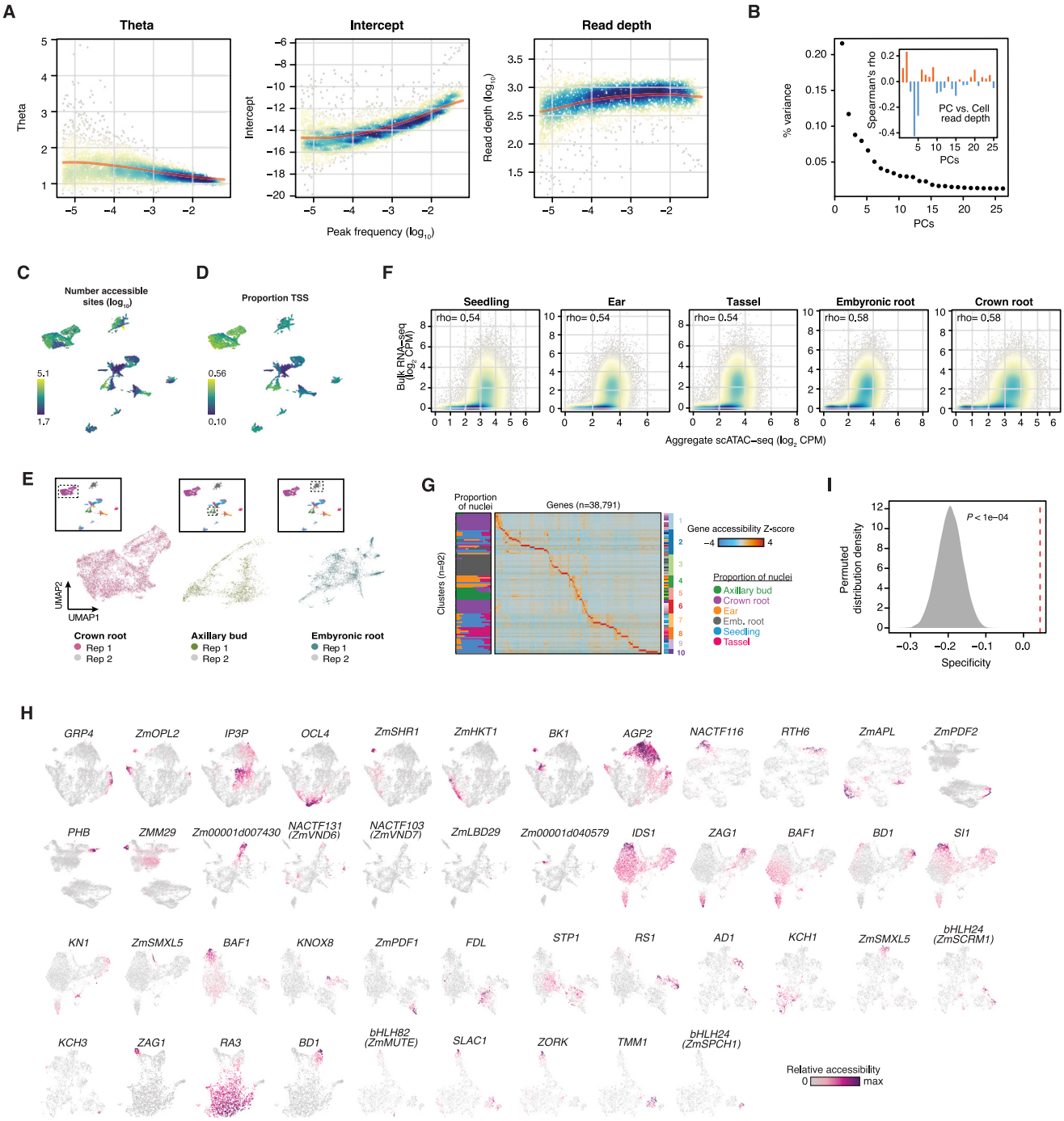


Figure S2. Clustering metrics and cell-type marker gene analysis, related to Figure 1

(A) Parameter regularization of model coefficients (y-axes) with respect to ACR usage (x-axes; proportion of nuclei with at least one Tn5 integration site in an ACR).

(B) Proportion of variance captured by the first 26 PCs. Inset: Spearman's correlation of principal components with cell read depth (log_{10}-transformed).

(C) Number of accessible sites per cell (log_{10}).

(D) Proportion of Tn5 integrations within 2-kb of gene TSSs per cell.

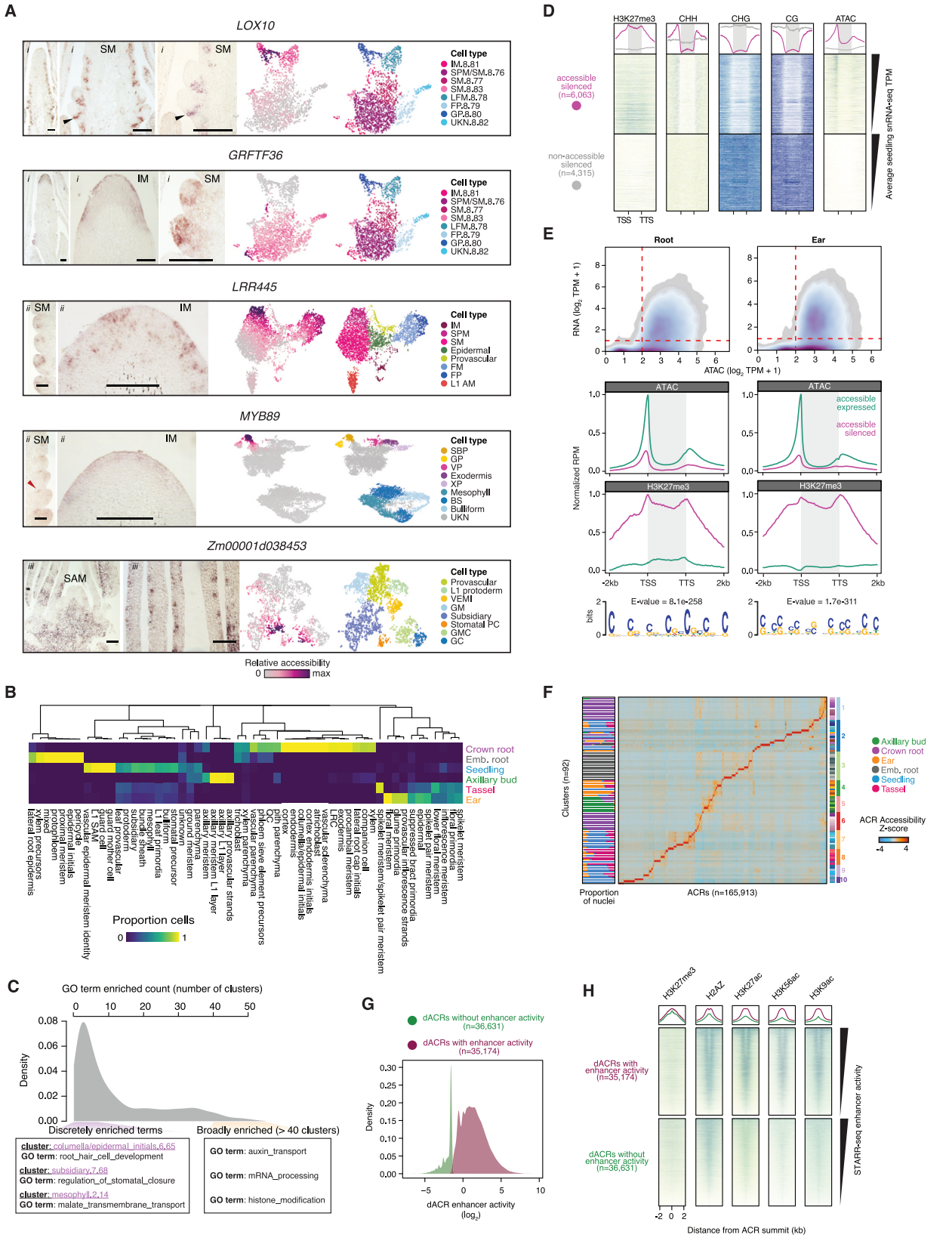
(E) Co-localization of nuclei barcodes from different biological replicates for three organs.

(F) Comparison of bulk RNA-seq expression levels (y axis, log_2 CPM) versus aggregate scATAC-seq gene accessibility scores (x axis, log_2 CPM) within an organ.

(G) Left: proportion of nuclei by organ in each sub-cluster. Right: gene accessibility (gene body plus 1-kb upstream TSSs) Z-scores across cell types.

(H) UMAP embeddings of nuclei barcodes colored by low (gray) to high (dark purple) gene accessibility scores (gene bodies plus 1-kb upstream TSSs) of cell type-specific marker genes.

(I) Permuted (10,000) distribution of average gene specificity scores for random sets of genes (gray) compared to the average specificity score of *a priori* marker genes (dashed red line).



(legend on next page)

Figure S3. Chromatin accessibility variation across plant cell types, related to Figure 3

(A) RNA *in situ* hybridization showing expression of *LOX10* in glume primordia and *GRFTF36* in IM and SMs of staminate inflorescence; *LRR445* in the IM periphery and SPMs and *MYB89* in the IM and suppressed bract primordia of pistillate inflorescence; Zm00001d038453 in ground tissue of SAM and leaf primordia sections. Gene accessibility scores and predicted cell types are shown on the right. *i*, tassel primordia. *ii,* ear primordia. *iii*, SAM/leaf. Black triangles point to the glume primordia. Red triangles point to suppressed bract primordia. Size bars illustrate 100-um. AM, axillary meristem; BS, bundle sheath; GC, guard cell; GM, ground meristem; GMC, guard mother cell; GP, glume primordia; IM, inflorescence meristem; L1, layer 1; LFM, lower floral meristem; SAM, shoot apical meristem; SBP, suppressed bract primordia; SM, spikelet meristem; SPM; spikelet pair meristem; Stomatal PC, stomatal precursor; UKN, unknown; VEMI, vascular/epidermal meristematic identity; VP, vascular parenchyma; XP, xylem parenchyma.

(B) Proportion of cells within subcluster (column) derived from one of six organs (rows).

(C) Distribution of GO term enrichment across clusters, where the x axis indicates the number of clusters in which a GO term is significantly enriched.

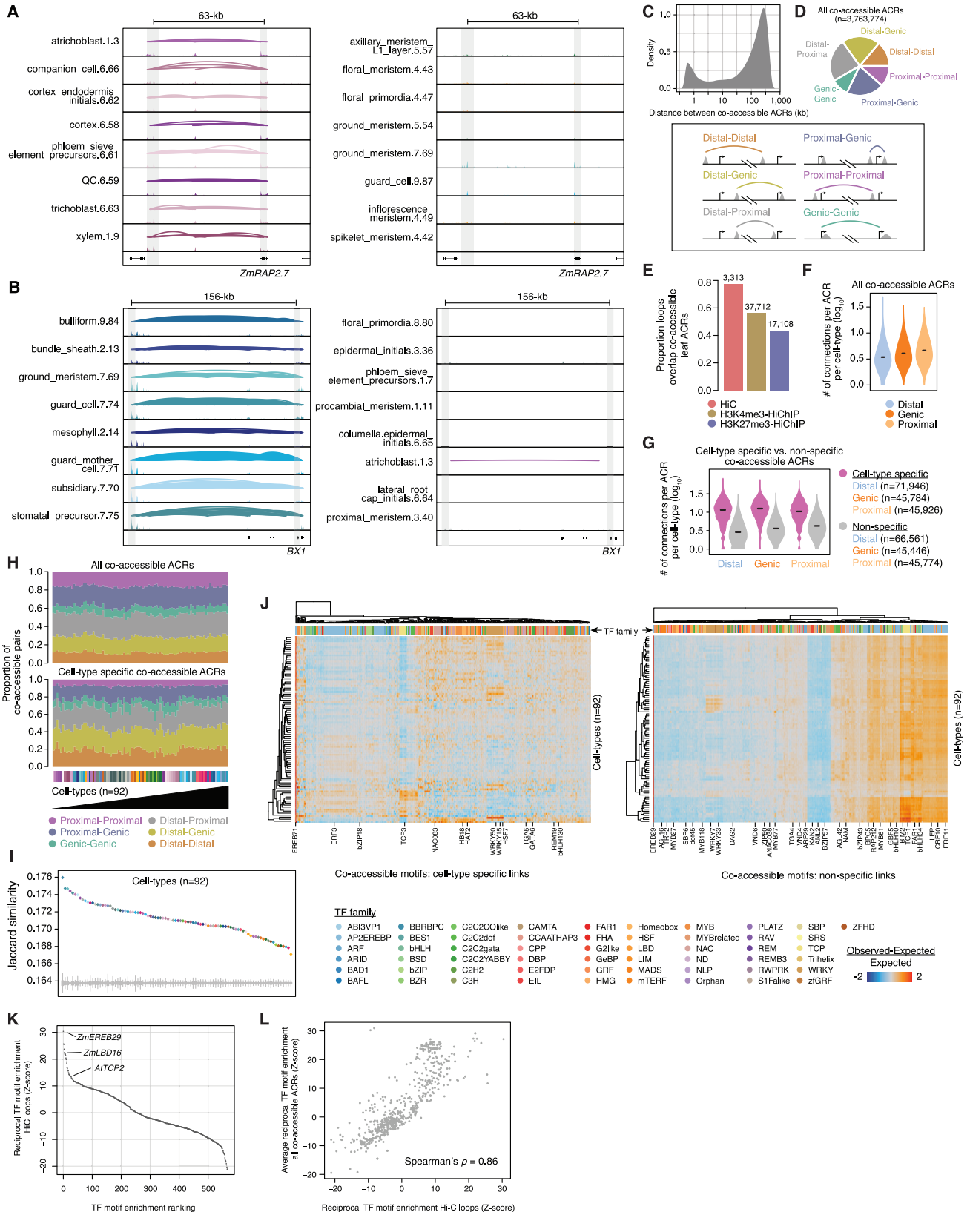
(D) Comparison of H3K27me3, mCHH, mCHG, mCG and chromatin accessibility between accessible/silenced genes (top heatmap, pink) and non-accessible/silenced genes (bottom heatmap, gray). Heatmaps were ordered by the average TPM values from snRNA-seq of maize seedlings.

(E) Top: Comparison of normalized ATAC-seq and RNA-seq reads in maize embryonic roots (left) and female inflorescence (ear; right). Horizonal and vertical dashed red lines indicate thresholds for accessibility (left: non-accessible, right: accessible) and RNA expression (below: non-expressed, above: expressed). Middle: Metaplots of ATAC-seq and H3K27me3 ChIP-seq reads 2-kb up and downstream of accessible/expressed and accessible/silenced genes from bulk root (left) and female inflorescence (ear; right) tissues. Bottom: Top de *novo* enriched motifs in ACRs within 1-kb upstream of TSSs for accessible/silenced genes in embryonic root (left) and female inflorescence (ear; right).

(F) Row ACR accessibility Z-scores across cell types. Cell types are ordered according to Table S2: Cluster Annotation and Metrics.

(G) Distribution of enhancer activity for distal ACRs (dACRs) classified as with (pink) and without (green) enhancer activity.

(H) Maize seedling ChIP-seq profiles for dACRs with (top heatmap, pink) and without (bottom heatmap, green) enhancer activity. Heatmaps were ordered by enhancer activity measured by STARR-seq of maize leaf protoplasts.



(legend on next page)

Figure S4. Co-accessible ACRs reflect *in vivo* chromatin interactions driven by coordinated TF activity, related to Figure 5

(A) Co-accessible ACRs at the *ZmRAP.2* locus in maize, with root-specific expression patterns, across eight root-derived (left) and eight above-ground (right) cell types. The height of the loops reflects the strength of co-accessibility. Pseudobulk chromatin accessibility tracks are shown under co-accessible ACR linkages for each cell-type.

(B) Co-accessible ACRs at the *BX1* locus in maize, predominantly expressed in seedling tissue, across eight seedling-derived (left) and eight non-seedling derived cell types. The height of the loops reflects the strength of co-accessibility. Pseudobulk chromatin accessibility tracks are shown under co-accessible ACR linkages for each cell-type.

(C) Distribution of physical distances between co-accessible ACR summits.

(D) Proportions of co-accessible ACR types, illustrated by toy examples.

(E) Proportion leaf Hi-C, H3K4me3-HiChIP and H3K27me3-HiChIP chromatin loops that overlap co-accessible ACRs from leaf cell types (clusters with greater than 50% of cells derived from seedlings).

(F) Log_{10} number of connections per ACR per cell type from all co-accessible ACRs split by genomic context: distal, proximal, and genic.

(G) log_{10} number of connections per ACR per cell type from cell type-specific (purple) and non-specific (gray) co-accessible ACRs, split by genomic context.

(H) Proportion of co-accessible classifications by cell type for all co-accessible ACRs (top) and cell type-specific co-accessible ACRs (bottom).

(I) Jaccard similarity of motif composition between co-accessible ACR edges by cell type (colored diamonds) relative to the same number of random ACR-ACR links, permuted 1,000 times for each cell-type (gray boxplots). Boxplots represent the interquartile range, gray lines indicate the permuted range.

(J) Heatmaps of observed proportion of co-accessible ACRs with the same motif embedded within link edges subtracted and divided by the expected proportion estimated by 1,000 permutations using sets of random ACR-ACR links.

(K) Ranked reciprocal TF enrichment in maize seedling Hi-C loops versus 1,000 permuted interactions containing a similar distribution of ACRs. Notable TF motifs are indicated by text.

(L) Comparison of reciprocal TF motif enrichment between maize seedling Hi-C loops and the average reciprocal TF motif enrichment from co-accessible ACRs across cell types.

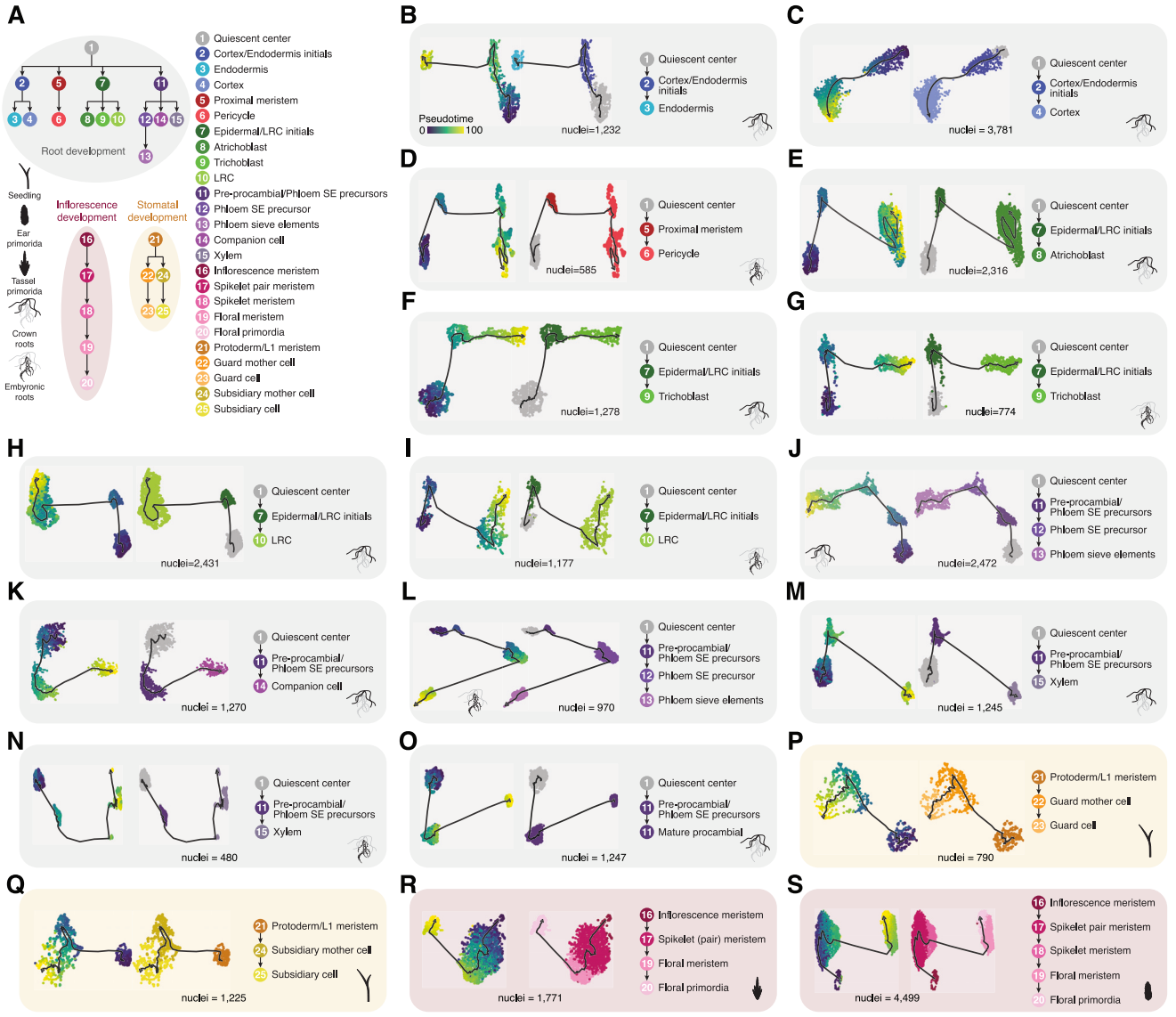


Figure S5. Pseudotime trajectory construction, related to Figure 6

(A) Overview of pseudotime developmental trajectory analysis from four organs: root, seedling, tassel (staminate inflorescence), and ear (pistillate inflorescence).
(B) Endodermis development in crown roots.
(C) Cortex development in crown roots.
(D) Pericycle development in embryonic roots.
(E) Atrichoblast development in crown roots.
(F) Trichoblast development in crown roots.
(G) Trichoblast development in embryonic roots.
(H) Lateral root cap (LRC) development in crown roots.
(I) Lateral root cap (LRC) development in embryonic roots.
(J) Phloem sieve element (SE) development in crown roots.
(K) Companion cell development in crown roots.
(L) Phloem sieve element (SE) development in embryonic roots.
(M) Xylem development in crown roots.
(N) Xylem development in embryonic roots.
(O) Procambial development in crown roots.
(P) Guard cell development in seedling.
(Q) Subsidiary cell development in seedlings.
(R) Floral primordia development in staminate inflorescence (tassel).
(S) Floral primordia development in pistillate inflorescence (ear).

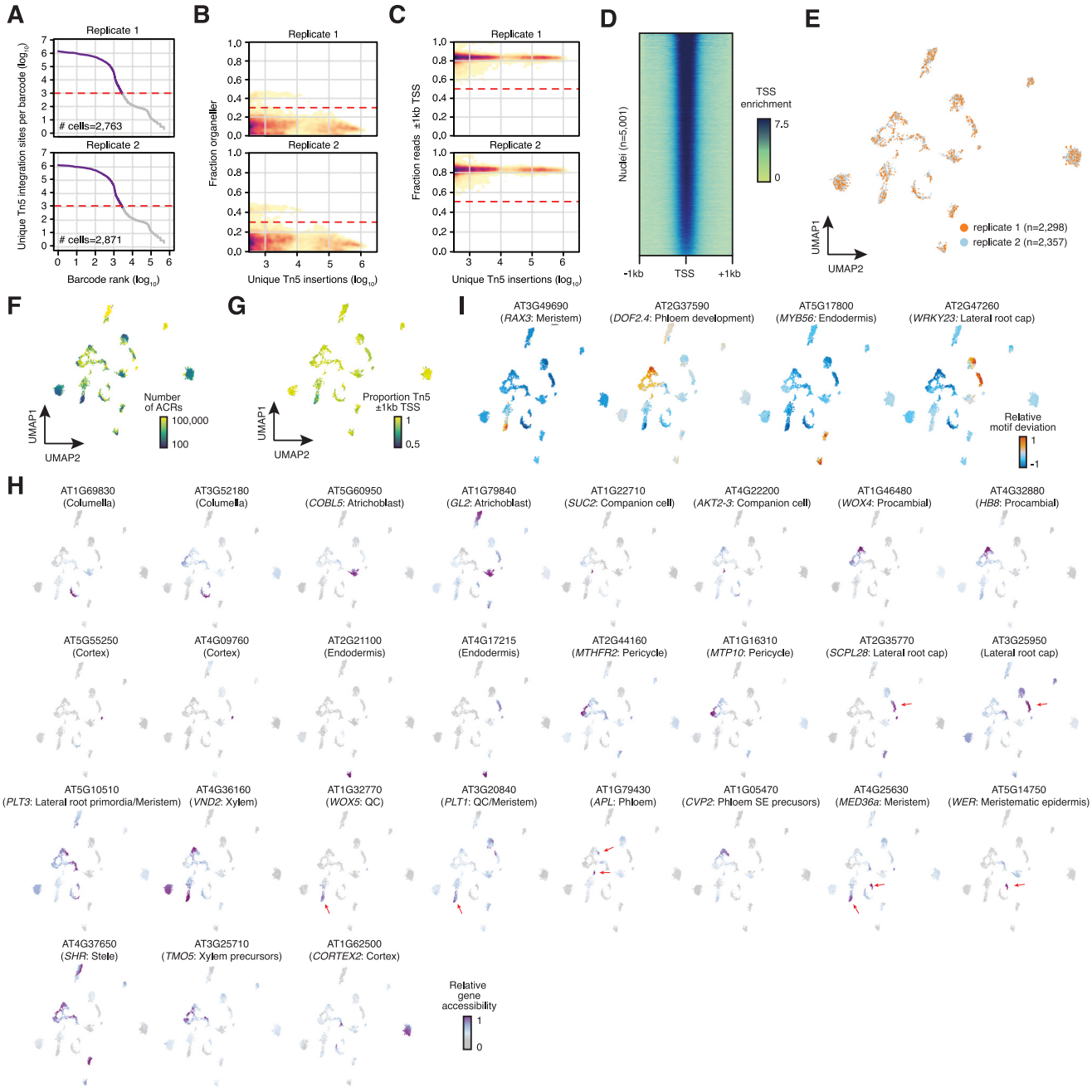


Figure S6. *Arabidopsis thaliana* root cell type atlas, related to Figure 7

(A) Knee plots for *Arabidopsis thaliana* root samples illustrating $\log_{10}$ transformed cellular read depths of $\log_{10}$ ranked barcodes across two biological replicates.

(B) Density scatterplots of $\log_{10}$ transformed barcode read depths (x axis) by the fraction of Tn5 integration sites derived from organellar sequences (chloroplast and mitochondrial) relative to the total number of unique Tn5 integration sites associated with each barcode from the two biological replicates. Dashed red lines indicate the threshold of two standard deviations from the mean used to filter lower quality barcodes.

(C) Density scatterplots of $\log_{10}$ transformed barcode read depths (x axis) by the fraction of Tn5 integration sites mapping to within 2-kb of transcription start sites (TSSs). Dashed red lines indicate the threshold of two standard deviations from the mean used to filter lower quality barcodes.

(D) Average TSS enrichment (normalized read depth adjusted by the two 10 bp windows 1-kb away from TSSs) across 5,001 *Arabidopsis thaliana* root barcodes (rows).

(E-G) UMAP (Uniform manifold approximation projection) embeddings of *Arabidopsis thaliana* root barcodes colored by (E) biological replicate, (F) the total number of accessible chromatin regions (ACRs), and (G) the proportion of Tn5 integration sites within 1-kb of TSSs.

(H) Relative gene accessibility for 27 known cell-type/domain restricted marker genes used to inform cell-type annotation of *Arabidopsis thaliana* root clusters.

(I) Relative motif deviations for transcription factors with known cell-type specificities.

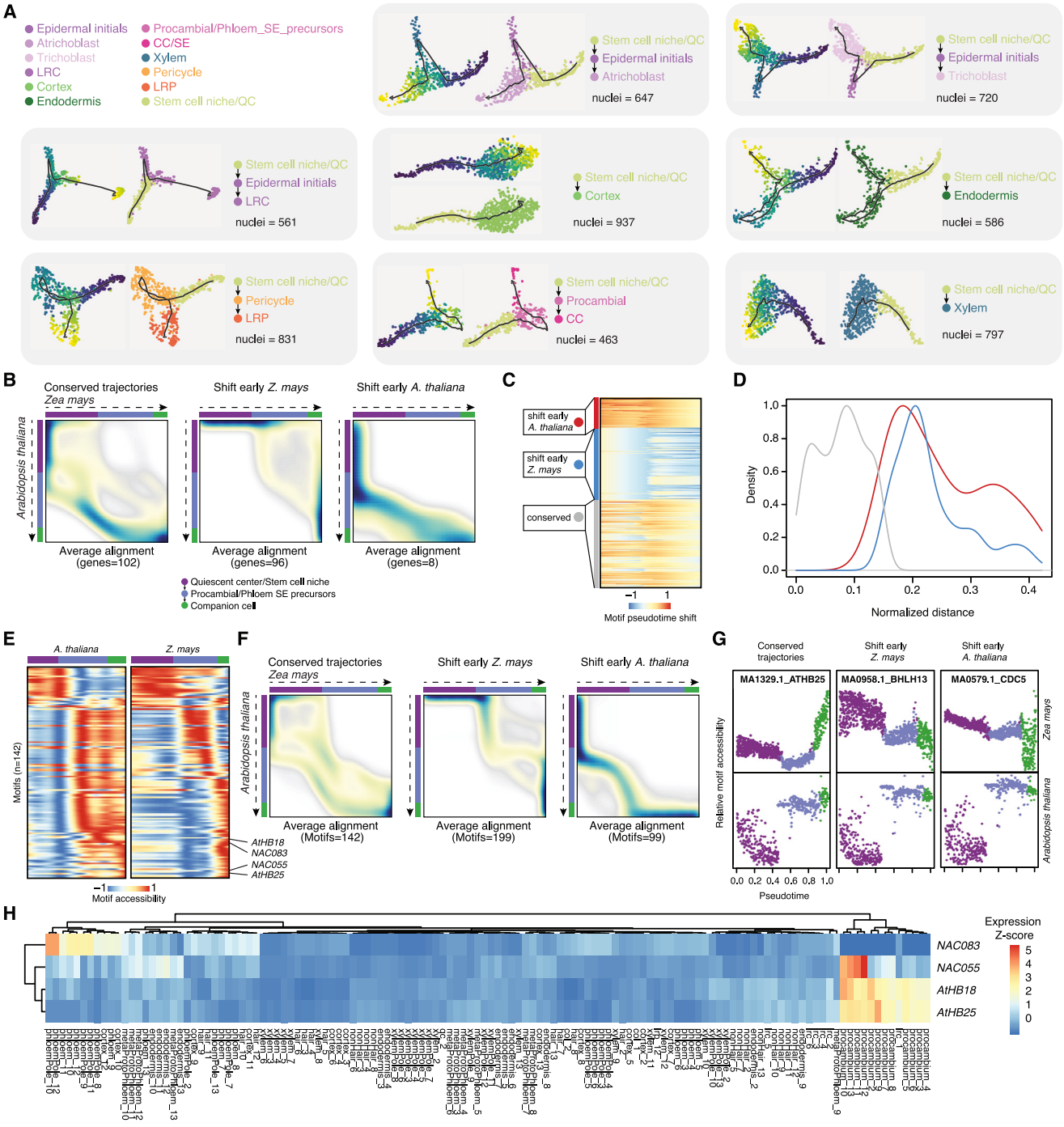


Figure S7. Dynamic and conserved chromatin accessibility across pseudotime between *Arabidopsis thaliana* and *Zea mays*, related to Figure 7

(A) Pseudotime trajectories for Atrichoblast, Trichoblast, Lateral root cap (LRC), Cortex, Endodermis, Lateral root primordia (LRP), Companion cells (CC), and Xylem development.

(B) Averaged alignments of conserved, shift early *Z. mays*, and shift early *A. thaliana* putative orthologs.

(C) Pseudotime shifts of TF motifs between *A. thaliana* and *Z. mays*, clustered into k-means and conserved groups.

(D) Distributions of motif-motif normalized distances between *Z. mays* and *A. thaliana* for the three groups.

(E) Conserved motifs (n = 142) ordered by pseudotime. Heatmaps for *A. thaliana* and *Z. mays* have identical row orders.

(F) Averaged alignments of conserved, shift early *Z. mays*, and shift early *A. thaliana* groups based on motif-motif global alignments from the dynamic time-warping algorithm.

(G) Examples of conserved, shift early *Z. mays*, and shift early *A. thaliana* motifs from both species.

(H) Gene expression Z-scores across *A. thaliana* FAC sorted root cell-types for the TFs recognizing the top four conserved motifs.