bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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]*

Affiliations:
[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

Correspondence: Alexandre P. Marand (marand@uga.edu), Robert J. Schmitz (schmitz@uga.edu)

ABSTRACT

***Cis*-regulatory elements (CREs) encode the genomic blueprints for coordinating spatiotemporal gene expression programs underlying highly specialized cell functions. To identify CREs underlying cell-type specification and developmental transitions, we implemented single-cell sequencing of Assay for Transposase Accessible Chromatin in an atlas of *Zea mays* organs. We describe 92 distinct states of chromatin accessibility across more than 165,913 putative CREs, 56,575 cells, and 52 known cell-types in maize using a novel implementation of regularized quasibinomial logistic regression. Cell states were largely determined by combinatorial accessibility of transcription factors (TFs) and their binding sites. A neural network revealed that cell identity could be accurately predicted (>0.94) solely based on TF binding site accessibility. Co-accessible chromatin recapitulated higher-order chromatin interactions, with distinct sets of TFs coordinating cell type-specific regulatory dynamics. Pseudotime reconstruction and alignment with *Arabidopsis thaliana* trajectories identified conserved TFs, associated motifs, and *cis*-regulatory regions specifying sequential developmental progressions. Cell-type specific accessible chromatin regions were enriched with phenotype-associated genetic variants and signatures of selection, revealing the major cell-types and putative CREs targeted by modern maize breeding. Collectively, our analysis affords a comprehensive framework for understanding cellular heterogeneity, evolution, and *cis*-regulatory grammar of cell-type specification in a major crop species.**

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

MAIN TEXT

Multicellular eukaryote development relies on precise spatiotemporal gene expression mediated by *cis*-regulatory elements (CREs)[1-7]. Despite the economic and nutritional importance of crops to society, our understanding of cell-type functions in plants has been precluded by technical limitations imposed by the cell wall and an inability to culture homogenous cell lines. To circumvent these limitations and comprehensively assess *cis*-regulatory variation among cell-types in a major crop species, we isolated nuclei using fluorescence-activated nuclei sorting (FANS) and generated single-cell accessibility profiles using Assay for Transposase Accessible Chromatin (scATAC-seq) from six major *Zea mays* L. cultivar B73 organs, including axillary buds, staminate and pistillate inflorescence, whole seedling (composed of stem, leaf, and coleoptile tissues), embryonic root tips, and post-embryonic crown roots (**Fig. 1a**, **Extended Data Fig. 1**, **Supplementary Table 1**). Rigorous evaluation of quality metrics, including doublet analysis via genotype mixing (B73/Mo17), biological replicates, per nucleus read-depth, transcription start site (TSS) enrichment, and fragment length distribution were reflective of high-quality scATAC-seq data (Methods, **Extended Data Fig. 1, 2**). In total, we identified 56,575 high-quality nuclei with an average of 31,660 unique Tn5 integration sites per nucleus (**Supplementary Data 1**). To group nuclei with similar global patterns of accessibility, we developed a regularized quasibinomial logistic regression framework that removes technical variation stemming from differences in nuclei read depth by pooling information across accessible chromatin regions (ACRs) and extracting model residuals as normalized accessibility scores for each nucleus and ACR (Methods). Denoising the residual matrix with singular value decomposition (SVD) and projecting the components into a two-dimensional space using Uniform Manifold Approximation Projection (UMAP) revealed 10 clusters generally composed of nuclei from the same organ, highlighting functional diversification of spatially distinct cells (**Fig. 1b, c; Extended Data Fig. 3a-e**). A second round of clustering for each major grouping uncovered a total of 92 sub-clusters (hereafter

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

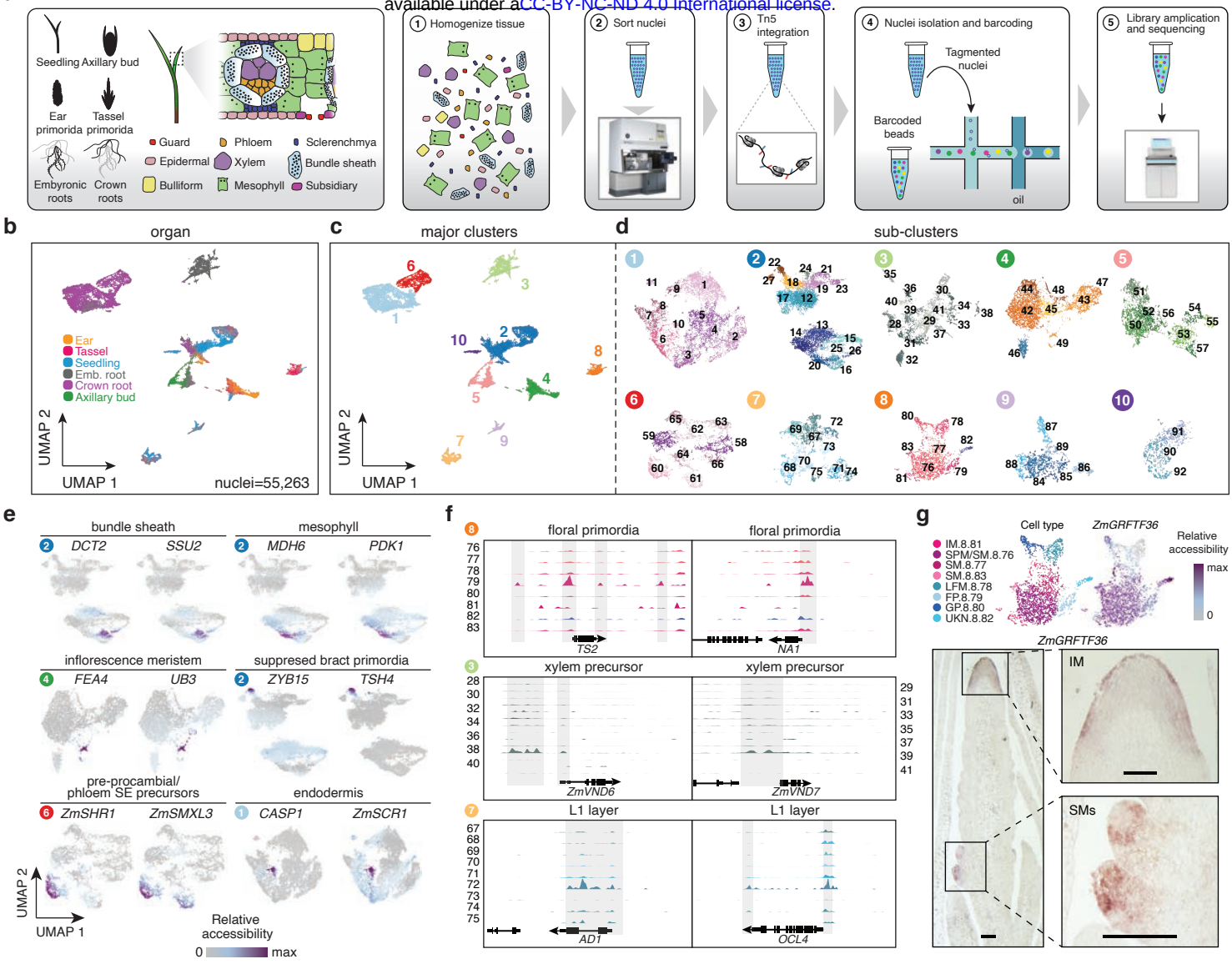


Figure 1: Atlas-scale profiling of single nuclei chromatin accessibility in *Zea mays*

a, Overview of experimental protocol. Homogenized samples release nuclei that are extracted by flow cytometry. Accessible chromatin is tagged by Tn5 and individual nuclei are distinctly barcoded in disparate droplets following microfluidic separation, allowing droplet pooling and library amplification prior to sequencing. **b,** Nuclei similarity clustering as a UMAP embedding derived from the denoised quasibinomial residuals across all ACRs for each nucleus. UMAP embedding of nuclei colored by **b,** organ identity and **c,** major cluster identity. **d,** UMAP embedding of sub-cluster assignments following a second round of clustering within each major cluster. **e,** Cell-type specific enrichment of gene accessibility for a subset of marker genes associated with six different cell-types. **f,** Sub-cluster specific chromatin accessibility profiles surrounding known marker genes for floral primordia, xylem precursors, and L1 epidermal cells. Bold, circled numbers indicate the cognate major cluster shown in panel **c**. Sub-cluster numeric identifications are present on the sides of the coverage plots. **g,** Top, gene accessibility for *ZmGRFTF36*, an inflorescence and spikelet meristem enriched transcription factor with no previously known cell-type specificity. Bottom, RNA *in situ* hybridization of *ZmGRFTF36* in maize B73 staminate (tassel) primordia. FP, floral primordia; GP, glume primordia; IM, inflorescence meristem; LFM, lower floral meristem; SM, spikelet meristem; SPM, spikelet pair meristem; UKN, unknown.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

referred to as clusters) with an average of 551 nuclei and 165,913 reproducible ACRs covering ~4% of the maize genome (**Fig. 1d**).

To annotate clusters with corresponding cell-types, we integrated gene and promoter (1-kb upstream TSSs) chromatin accessibility information with long-range distal ACRs on the basis of co-accessibility (Methods) into a single accessibility score for each gene and nucleus as a proxy for gene expression (bulk RNA-seq versus aggregate scATAC-seq Spearman's correlation coefficient = 0.54-0.58, **Extended Data Fig. 3f**). We then (*i*) evaluated differential accessibility among clusters for a manually curated list of 221 known marker genes (**Supplementary Table 2**), (*ii*) classified cell-types of individual nuclei with a multinomial logistic classifier trained on nuclei with discriminative cell-type specific signatures, and visually assessed (*iii*) accessibility scores of *a priori* marker genes over the UMAP embeddings and (*iv*) cluster-aggregated ATAC-seq coverages (Methods, **Fig. 1e**,**f, Extended Data Fig. 4a**, **Supplementary Table 3**). Patterns of gene accessibility were consistent with *a priori* information of cell-type/domain specific expression, such as overlapping accessibility of bundle sheath-specific genes *DICARBOXYLIC ACID TRANSPORTER1* (*DCT2*) and *RIBULOSE BISPHOSPHATE CARBOXYLASE SMALL SUBUNIT2* (*SSU2*) in bundle sheath cells, and mesophyll-specific genes *MALATE DEHYDROGENASE6* (*MDH6*) and *PYRUVATE DEHYDOGENASE KINASE1* (*PDK1*) in mesophyll cells (**Fig. 1e, f, Extended Data Fig. 4a**)[8]. To further 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, expression patterns were in line with the predicted localization based on gene accessibility (**Fig. 1g**, **Extended Data Fig. 4b**). Estimates of cell-type proportions within and across organs were also concordant with prior observations, such as the occurrence of vascular bundle sheath and parenchymal mesophyll cells within multiple organs, including those derived from stem and leaf tissues of seedlings and within pistilate and staminate inflorescence (**Extended Data Fig. 4c**, **Supplementary Table 3**)[9].

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Analysis of chromatin accessibility variation at genic loci revealed 74% (28,625/38,752) of genes with significant differential accessibility (FDR < 0.05, log fold-change > 2), with an average of 2,768 differentially accessible genes per cluster (**Extended Data Fig. 5b, Supplementary Data 2**). Marker-agnostic gene set enrichment analysis (GSEA) of Gene Ontology (GO) terms exemplified prior information regarding specific cell-type functions, with enriched terms such as "root hair cell development" in root epidermal initials, "regulation of stomatal closure" in subsidiary cells, and "malate transmembrane transport" for mesophyll cells (**Extended Data Fig. 4d**). Distinct cell-types were generalized by highly specific GO annotations, as most (>51%) GO terms were identified in only a handful of cell-types (five or fewer), implicating accessibility dynamics as underling the signature hallmarks of cell-type identity and function (**Extended Data Fig. 4d**). In summary, this combinatorial approach enabled the identification of 52 cell-types for 83% (76/92) of clusters, capturing nearly all major expected cell-types in the profiled organs and suggesting the existence of novel uncharacterized cell-types present in these data (**Supplementary Table 3**).

Characterization of chromatin accessibility variation

Deconvolution of nuclei into discrete cell-types provides an opportunity to identify CREs encoding cell identity. To this end, we implemented a generalized linear model to catalog ACRs with discrete patterns of chromatin accessibility across cell-types (Methods). In total, 52,520 ACRs (31%) were differentially accessible (FDR < 0.05, fold-change > 2) and restricted to one or a handful of clusters, with an average of 2,826 per-cluster (**Extended Data Fig. 5a**, **6a**, **b**). Similar to the reported functions of transcriptional enhancers in metazoan genomes, we found cluster-specific ACRs were associated with significantly greater enhancer activity based on STARR-seq of maize leaf protoplasts[10] relative to controls (Methods, n=165,913) and non-specific (n=113,393) ACRs (**Fig. 2a**). Deconvolution of chromatin accessibility by cell-type revealed accessible sites primarily located distal to genic regions (>2-kb from any gene) compared to previously published

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

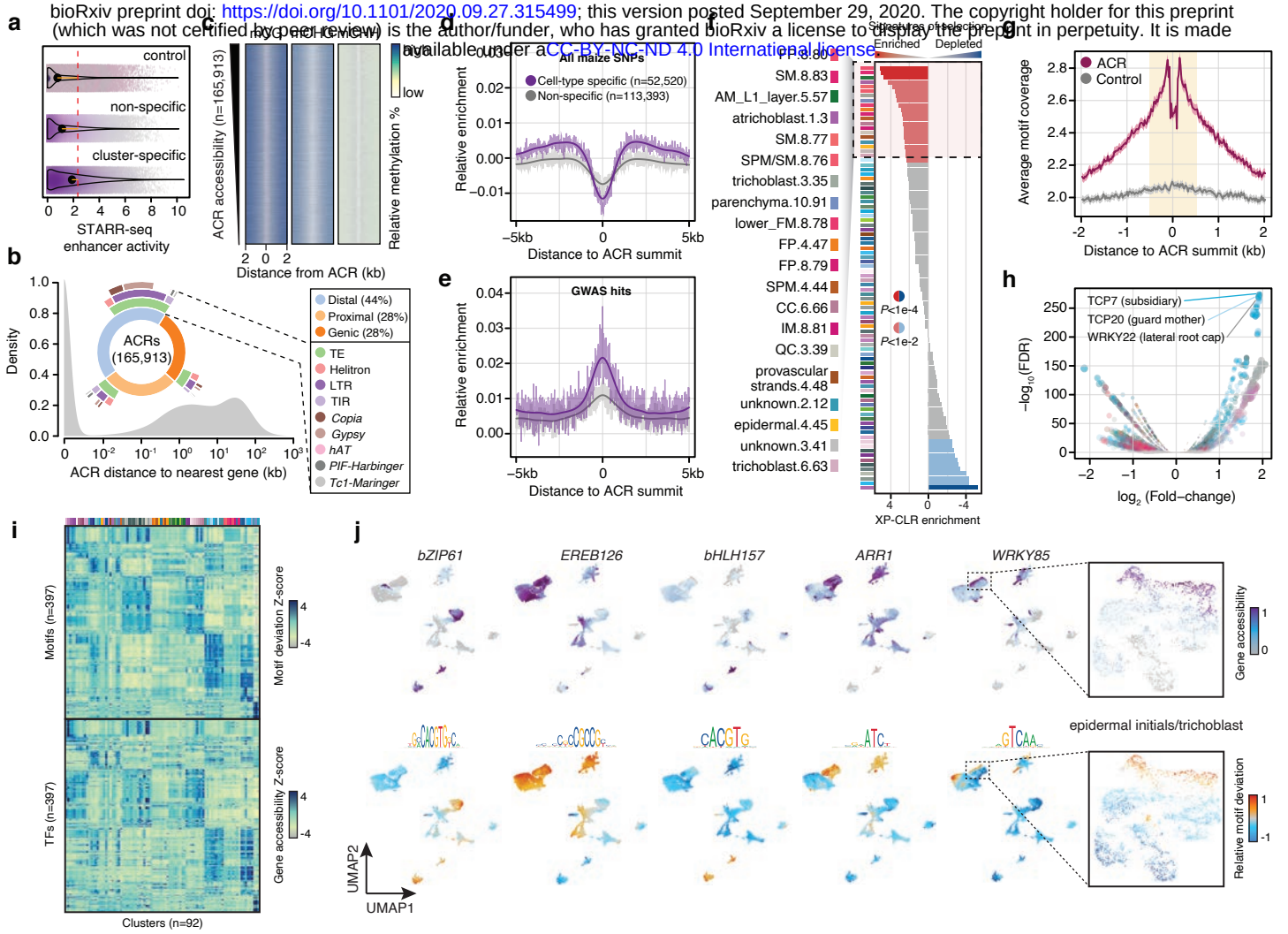


Figure 2: Characterization of accessible chromatin regions

a, Distribution of enhancer activity (maximum log2[RNA/input]) for control regions (n=165,913), non-specific (n=113,393) and cluster-specific ACRs (n=52,520). The dash red line indicates the overall mean. Orange lines reflect differences between the group mean and overall mean. **b,** Distribution of ACR distances to the nearest gene. Inset, distribution of ACR genomic context. **c,** Relative DNA methylation levels 2-kb flanking ACRs. **d,** Relative enrichment of polymorphisms after normalizing by mappability for 5-kb regions flanking cell-type specific and non-specific ACR summits. Smoothed splines are shown as dark lines. **e,** Relative enrichment of significant GWAS polymorphisms relative to all polymorphisms for 5-kb regions flanking cell-type specific and non-specific ACR summits. **f,** Enrichment of signatures of selection (XP-CLR) in the top 2,000 ACRs for all cell-type clusters. The 20 most enriched cell-types are highlighted on the left. AM, axillary meristem; CC, companion cell; FM, floral meristem; FP, floral primordia; IM, inflorescence meristem; QC, quiescent center; SM, spikelet meristem; SPM, spikelet pair meristem. **g,** Average motif coverage for all ACRs (n=165,913) and control regions (n=165,913). Shaded polygon, 95% confidence intervals. **h,** Enrichment of TF motifs in the top 2,000 ACRs ranked by Z-score for each cell-type compared to the top 2,000 most constitutive ACRs via binomial tests. FDR was estimated by the Benjamini-Hochberg method. **i,** Comparative heatmaps for matched TF gene accessibility (bottom) and motif deviation (top) Z-scores across clusters. **j,** Gene accessibility scores for five maize transcription factors (top) and their associated motif deviations (bottom).

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

bulk-level experiments (**Fig. 2b**)[10]. Notably, 30% (22,456/73,791) of distal ACRs overlapped with LTR transposons, including the major maize domestication locus *tb1*-enhancer, and were generally devoid of DNA methylation[11,12]. These findings are consistent with transposable elements playing a prominent role in CRE evolution of the maize genome (**Fig. 2b, c, Extended Data Fig. 1a**)[13-15].

Sequence variation underlying CREs contribute to disease emergence and phenotypic innovation over evolutionary timescales[16,17]. In contrast to broadly accessible regions, analysis of extant genetic variation in maize revealed lower polymorphism rates within cell-type specific ACRs (**Fig. 2d**). However, of the genetic variants embedded within ACRs, those within cell-type specific ACRs were more frequently associated with phenotypic variation[18] (**Fig. 2e**). To investigate the contribution of domestication and signatures of selection in distinct cell-type contexts, we assessed the relative enrichment of multi-locus allele-frequency differentiation signal (XP-CLR scans) from chronologically sampled elite inbred maize lines within cell-type specific ACRs[19]. Of the 21 cell-types with significant (FDR < 0.01) selection signature enrichment, 57% (12) correspond to staminate and pistillate cell-types, such as spikelet meristems, spikelet pair meristems, inflorescence meristems, floral meristems, and floral primordia (**Fig. 2f**). Thus, modern maize breeding resulted in the selection of alleles containing floral-specific ACRs associated with agronomically favorable inflorescence architecture[20].

To characterize the determinants of chromatin accessibility variation among cell-types, we mapped the genome-wide occurrences of 570 transcription factor (TF) motifs. Expectedly, ACRs were highly enriched with putative TF binding sites relative to control (n=165,913) and flanking regions (**Fig. 2g**). Using the top 2,000 differential ACRs for each cell-type, we found 84% of TF motifs (475/568) were enriched (binomial test: FDR < 0.05, Methods) in at least one type, with a median of 43 enriched TF motifs per cell-type (**Fig. 2h**). To characterize TF motif accessibility variation at the scale of individual nuclei, we estimated relative TF motif enrichment controlling for background and sequence bias across ACRs for each nucleus (**Extended Data Fig. 6c-f**).

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

153 Diverse combinatorial TF motif landscapes were apparent across cell-types and closely mirrored
154 gene accessibility profiles from matched TFs (median PCC across cell-types = 0.45), establishing
155 synchronized chromatin accessibility of *cis* and *trans* cell autonomous factors as major
156 determinants of cell identity (**Fig. 2i, j**, **Extended Data Fig. 6g**). To benchmark the utility of TF
157 motifs for explaining cell identities, we trained a neural network (NN) to learn patterns of TF motif
158 accessibility underlying various cell-types. The NN model achieved an overall accuracy of 0.94
159 and an average sensitivity and specificity of 0.93 and 0.99, respectively, indicating that patterns
160 of motif accessibility enable highly predictive classifications of diverse cell states (**Extended Data**
161 **Fig. 6h**). Furthermore, assessment of enriched TFs and their cognate motifs identified several
162 known cell-type specific regulators – including *WRKY* family TFs in root epidermal progenitors
163 and trichoblasts[21], *G2-like1* in parenchymal mesophyll[8], and *AGAMOUS-like* and *SEPALLATA*[22]
164 TFs in floral primordia – as well as previously uncharacterized TFs with new potential roles as
165 cell-type regulators (**Fig. 2i, j**, **Supplementary Table 4, 5**). Taken together, these results are
166 indicative of robust inference of CRE and TF activity at the level of single cells and reveal TF
167 dynamics central to *cis*-regulatory specification of diverse cell states.
168
169 **Coordinated dynamic chromatin accessibility recapitulates *in vivo* chromatin interactions**
170 Correlated changes in chromatin accessibility of nearby loci represent putative physical chromatin
171 interactions with regulatory potential[1,23-26]. We identified 3.8 million (M) ACR-ACR linkages
172 (hereafter referred to as co-accessible ACRs) with significantly correlated (empirical FDR < 0.05,
173 Methods) patterns of chromatin accessibility across cell-types, capturing known physical
174 interactions such as between *TB1* and the genetically mapped domestication *tb1*-enhancer that
175 confers inflorescence/axillary meristem-specific expression (**Fig. 3a**, **Extended Data Fig.**
176 **7a**)[10,14,27,28]. To assess the broad interactive potential of co-accessible ACRs *in vivo*, we compared
177 co-accessible ACRs from leaf cell-types with maize whole-leaf chromatin conformation capture
178 data, recovering more than 78% (3,313/4,265), 57% (37,712/65,691) and 44% (17,108/38,567)

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

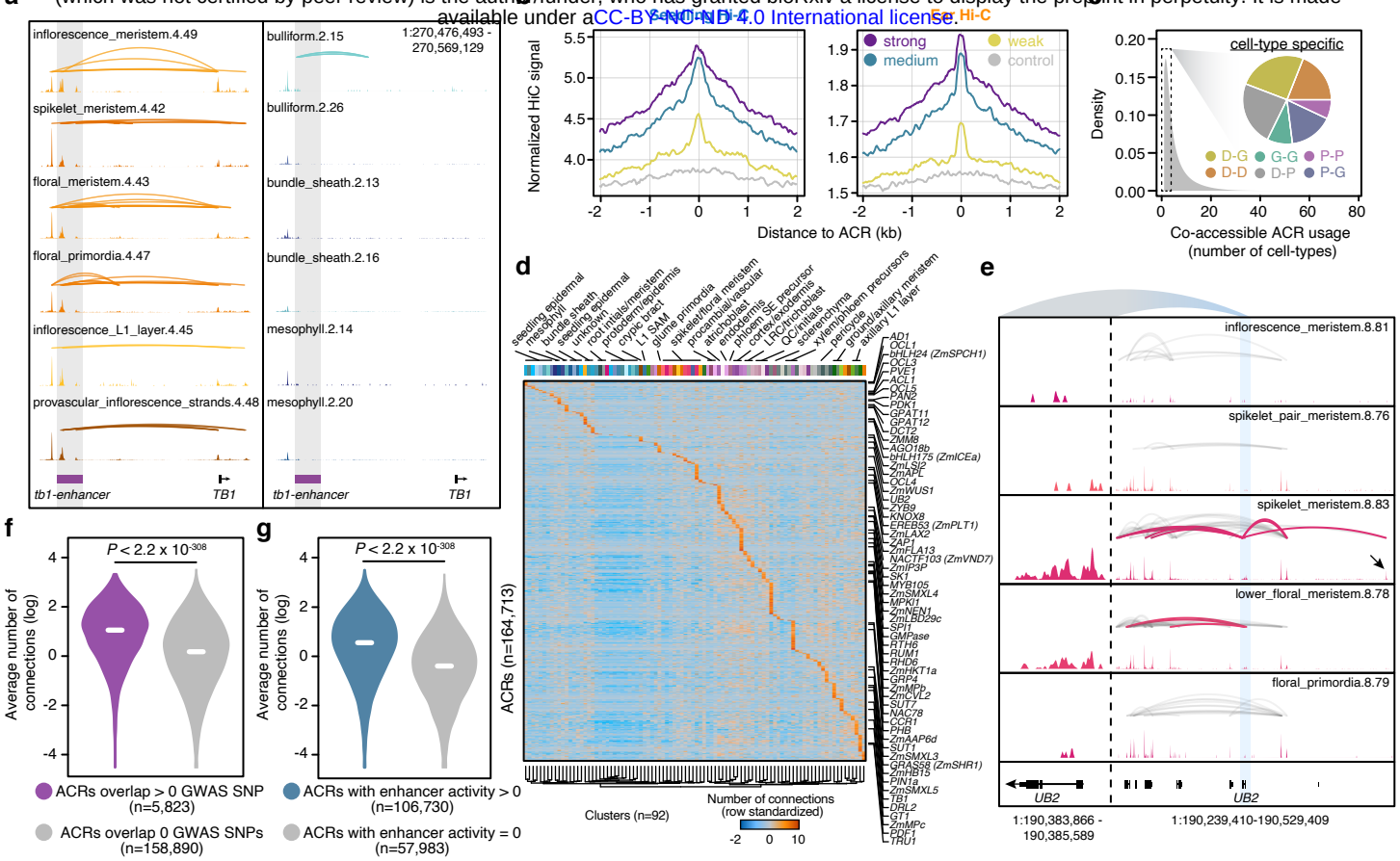


Figure 3: Co-accessible ACRs reflect *in vivo* chromatin interactions and are established by cell-type specific TFs. **a**, Comparison of ear and seedling co-accessible ACRs across nine cell-types at the genetically mapped *tb1-enhancer* domestication locus (highlighted region). Link height reflects co-accessibility scores between ACRs across cells in a cluster. **b**, Average normalized Hi-C signal across 4-kb windows centered on ACRs equally distributed into three groups (each group: n=54,904) based on the number and strength of participating co-accessible links. **c**, Co-accessible ACR interaction frequency across clusters. Inset: Single cell-type co-accessible links (n=1,018,417) for different genomic contexts. **d**, Standardized (row-wise) number of connections for each ACR (rows) by cell-type (columns). Column color map reflect cell-types from the legend in Extended Data Fig. 4. Gene-proximal ACRs for a subset of marker genes are indicated on the right. **e**, Right, chromatin accessibility and co-accessible ACR links surrounding the *UB2* locus associated with ear row number and tassel branch number quantitative traits. Black arrow indicates a distal ACR upstream of *UB2* present only in spikelet meristems. Co-accessible links with an edge within 2-kb of *UB2* are colored pink while remaining links are grey. Link height represents co-accessibility strength. Left, close-up of accessibility profiles of *UB2*. **f**, Distributions of average ACR-ACR links across cell-types for ACRs that overlap (purple) and do not overlap (grey) phenotype-associated genetic variants from maize GWAS. The median of each distribution is shown as a white horizontal line. Violin plots present the entire range of average number of connections on a log scale. Hypothesis testing was conducted within the *R* statistical framework via Wilcoxon rank sum test. **g,** Distributions of average ACR-ACR links celltypes for ACRs with (blue) and without (grey) enhancer activity (log2 RNA/input greater than 0). Hypothesis testing and distribution illustration was performed similarly as panel **f**.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of chromatin loops from Hi-C, H3K4me3-HiChIP and H3K27me3-HiChIP experiments, respectively (**Extended Data Fig. 7b**). Hi-C/HiChIP is a direct reflection of the proportion of cells exhibiting a particular interaction, as ubiquitously interacting loci dominate rarer cell-type specific contexts that are frequently missed by loop-calling algorithms. To determine the relative predictability of *in vivo* interactions using co-accessible ACRs, we estimated the interaction strength of each ACR by integrating correlative scores across cell-types (Methods). ACRs classified by interaction strength recapitulated expected *in vivo* chromatin interaction frequencies, where even the weakest class of co-accessible ACRs were associated with elevated interaction frequencies relative to flanking and randomized control regions (**Fig. 3b**).

Cataloguing the usage of co-accessible ACRs across cell-types identified more than 27% (~1M in total) that were unique to a single type, 49% of which were classified as distal-genic/proximal, with an average of 11,069 distinct links per cell-type (**Fig. 3c**). Consistent with regulatory models where a single gene can interact with multiple distant loci, proximal ACRs, rather than distal or genic ACRs, were associated with the greatest number of links on average (Wilcoxon rank sum test: $P < 2.2e\text{-}308$, **Extended Data Fig. 7c**). Highlighting long-range "hub" interactions as a key contributors towards cell identity, cell-type specific co-accessible ACRs were associated with greater number of links per site (Wilcoxon rank sum test: $P < 2.2e\text{-}308$) and a greater proportion of links involving distal ACRs (Chi-squared test: $P < 2.2e\text{-}308$, **Extended Data Fig. 7c, d**). Furthermore, the interactive capacity of any given ACR strongly depended on the cell-type context (**Fig. 3d**). For example, *UNBRANCHED2* (*UB2*) – a major ear row number and tassel branch number quantitative trait locus[29] – demonstrated preferential accessibility in spikelet meristems that coincided with the greatest number of *UB2* proximal to distal ACR interactions, including a cell-type specific ACR located upstream approximately 150-kb (**Fig. 3e**). We posited that ACRs with expanded interactive capacity resemble enhancers with the potential to influence organismal phenotypes. Indeed, ACRs with enhancer activity and co-localization with phenotype-associated genetic variants from GWAS were associated with a significantly greater number of

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ACR-ACR connections (Wilcoxon rank sum; P < 2.2e-308, **Fig. 3f**, **g**). These results highlight the occurrence of diverse cell-type specific regulatory configurations among distal enhancer ACRs and their target genes and implicate genetic variants perturbing highly interactive distal enhancers as major contributors towards phenotypic variation.

The structural protein, CCCTC-binding factor (CTCF) plays an important role in metazoan genome organization and is notably absent in plant lineages[30]. In search of an orthogonal factor in maize, we hypothesized that higher-order chromatin structure captured by co-accessible ACRs may be driven by TFs recognizing similar sequence motifs embedded within interacting accessible regions. Comparison of motif occurrences between co-accessible link edges indicated that ACRs in co-accessible links are more similar to one another than randomly linked ACRs (empirical; P < 1e-4, **Extended Data Fig. 7e**). Furthermore, we identified several cases where co-accessible ACR edges were reciprocally enriched (FDR < 0.05, Methods) for the same TF motif in both cell-type specific and non-specific co-accessible links (**Extended Data Fig. 7f**). We identified *TEOSINTE BRANCHED1/CYCLOIDEA/PROLIFERATING CELL NUCLEAR ANTIGEN FACTOR1* (*TCP*), *APETALA2*/*ETHYLENE-RESPONSIVE ELEMENT BINDING PROTEINS* (*AP2-EREBP*) and *LATERAL ORGAN BOUNDARIES DOMAIN* (*LBD*) motifs that were not only broadly associated with co-accessible edges across cell-types (FDR < 0.05), but also exhibited strikingly similar GC-rich palindromic binding sites (**Extended Data Fig. 7f, g**). A role in chromatin organization is supported by previous research demonstrating *TCP* motif overrepresentation in topologically associated domain-like (TAD) boundaries in *Oryza sativa* and *Marchantia polymorpha,* and the distal edges of chromatin loops in *Zea mays*[27,28,31,32].

Dynamic chromatin accessibility specifies cell developmental trajectories

The apical domains of maize enclose undifferentiated meristematic stem cells that continuously give rise to all differentiated cell-types[33,34]. To define a *cis*-regulatory catalog of temporal cell-type progressions, we estimated pseudotime scores for ordered nuclei across 18 developmental

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

trajectories reflecting meristematic to differentiated cell-states; identifying ACRs, TF loci, and TF motifs with significant accessibility variation across pseudotime in a given trajectory (**Fig. 4a**, **Extended Data Fig. 8**, **Supplementary Data 3-6**, Methods). To showcase the power of trajectory construction to characterize a relatively understudied process, we focused our analysis on root phloem companion cell (PCC) development (**Fig. 4b**). Ordering nuclei along the PCC trajectory, we identified 8,004 ACRs, 440 TF motifs, 7,955 genes, and 402 TF loci that were differentially accessible across pseudotime (**Fig. 4c**, Methods). Several known meristem and phloem developmental genes including *AT-RICH INTERACTIVE DOMAIN-CONTAINING 8* (*ARID8*)[35], *SUPPRESSOR OF MAX2 1-LIKE3* (*ZmSMXL3*)[36], *and SUCROSE TRANSPORTER 1* (*ZmSUT1*)[37], were identified among the top differentially accessible genes throughout PCC development (**Fig. 4d**). To realize conserved regulatory mechanisms in plant root development, we profiled chromatin accessibility in 4,655 nuclei from 7-day old embryonic root tissues in the eudicot model species *Arabidopsis thaliana* and constructed eight pseudotime trajectories encompassing vascular, dermal and ground development (**Fig. 4e, f, Extended Data Fig. 9, Supplementary Data 7-11**). Maintaining focus on PCC development, we aligned *Z. mays* and *A. thaliana* trajectories using a time-warping algorithm. Consistent with recent comparative analysis of vascular development in *O. sativa*, *A. thaliana,* and *Solanum lycopersicum*[38], only 206 out of 10,976 putative orthologs were significantly associated (FDR < 0.01) with PCC pseudotime in both species, indicating that the majority PCC trajectories-associated genes are unique to each lineage (97% *Z. mays*, 83% *A. thaliana*). However, of the 206 PCC-associated orthologs, ~50% (102/206) exhibited similar patterns of gene accessibility across pseudotime (**Fig. 4g**, **Extended Data Fig. 10a, b**). Several putative orthologs with matching gene accessibility patterns have been previously associated with PCC development, such as *SCARECROW-LIKE8* (*SCL8*), which displays an increasing expression gradient in nascent to mature *A. thaliana* PCCs in a previously generated *A. thaliana* root cell-type gene expression atlas (**Fig. 4h**)[39]. Despite evolutionary

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

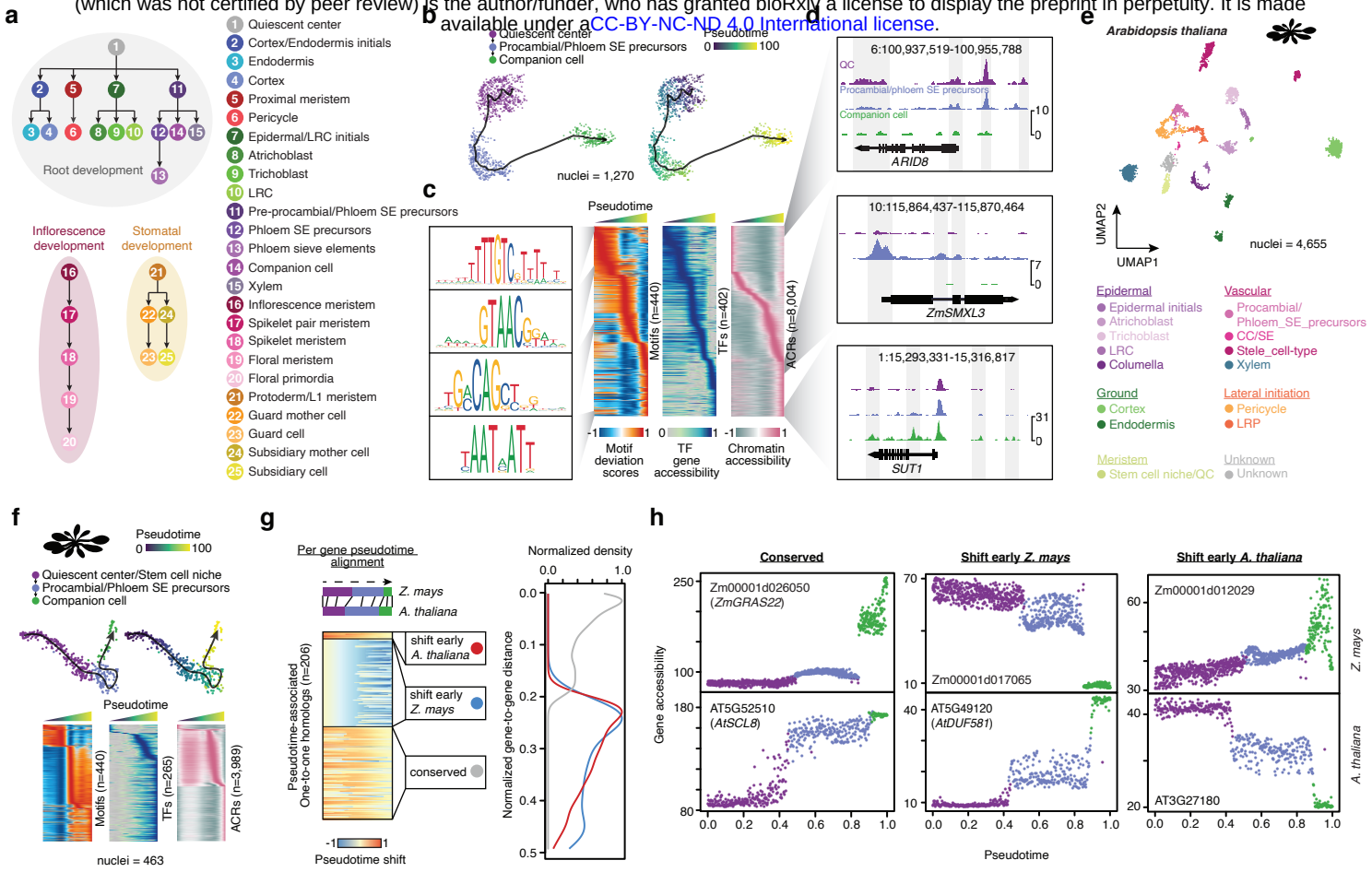


Figure 4: Chromatin accessibility is dynamic across pseudotime and conserved between *Arabidopsis thaliana* and *Zea mays*

a, Overview of pseudotime trajectory analysis. Inflorescence development was performed for both pistillate and staminate inflorescence. **b,** UMAP embedding of companion cell developmental trajectory depicting cell-types (left) and pseudotime progression (right). **c,** Relative motif deviations for 440 TF motifs (left, rows), 402 TF gene accessibility scores (middle, rows), and relative accessibility of 8,094 ACRs (right, rows) associated with pseudotime (columns). Four motifs enriched along the trajectory gradient are shown on the left. ACRs, accessible chromatin regions; TF, transcription factor. **d,** Genome browser screenshots of cell-type specific chromatin accessibility profiles along the developmental trajectory for quiescent center, procambial/phloem sieve element precursor, and companion cell at associated marker gene loci. **e,** UMAP embedding based on whole-genome chromatin accessibility profiles of 4,655 *Arabidopsis thaliana* root nuclei. **f,** UMAP embedding of compnaion cell developmental trajectories in *Arabidopsis thaliana* depecting cell-types (left) and pseudotime progression (right). Motifs, TFs, and ACRs with significant association with pseudotime are shown as heatmaps below. **g,** Per-gene pseudotime shift scores from alignments between *Zea mays* and *Arabidopsis thaliana* companion cell development progressions, clustered by k-means into three groups. Distribution of gene-gene distances from the alignments are shown on the right, split by k-mean groups. **h,** Exemplary one-to-one homologs between *Arabidopsis thaliana* and *Zea mays* for the three groups split by pseudotime shifts. Acronyms: CC, compansion cell; LRC, lateral root cap; LRP, lateral root primordia. QC, quiescent center; SE, sieve elements.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

conserved sequence homology, the remaining homologs (n=104) exhibited differential accessibility patterns clustered into two groups that reflect changes in the timing of gene accessibility across PCC development, underscoring functional novelty in PCC development between *Z. mays* and *A. thaliana*. To understand the extent of innovation in *cis* regulation along the pseudotime continuum, we aligned *Z. mays* and *A. thaliana* TF motif accessibility profiles associated with PCC progression (**Extended Data Fig. 10c**). Of the 440 motifs, 142 demonstrated highly conserved *cis*-regulatory dynamics between species (**Extended Data Fig. 10d, e**). Indeed, TFs recognizing the top four motifs ranked by normalized distances (Methods) included *HOMEOBOX25* (*HB25*), *HOMEOBOX18* (*HB18*), *NAC DOMAIN CONTAINING PROTEIN 55* (*NAC055*), and *NAC DOMAIN CONTAINING PROTEIN 83* (*NAC083*) that have been previously implicated in regulation of hormonal responses and vascular development[40-42]. Gene expression profiles of these TFs from published root cell-type resolved data in *A. thaliana* was largely restricted to maturing procambial and companion cells, consistent with motif accessibility dynamics in both *Z. mays* and *A. thaliana* ontogenies (**Extended Data Fig. 10f-h**)[39].

In summary, we describe 92 distinct cell-states from a suite of *Z. mays* organs, many of which previously lacked prior genome-wide characterization. We provide evidence supported by RNA *in situ* hybridization for the use of chromatin accessibility as a robust proxy of gene expression. We defined the TFs, CREs, and other genomic regions that discretize cell-type identities and the sequential trajectories for an array of developmental ontogenies. We demonstrate that floral cell-type specific ACRs have been the historical targets of modern agronomic selection and while subject to considerable sequence constraint within maize, cell-type specific ACRs harbor polymorphisms associated with significant phenotypic variation. By considering patterns of gene and motif accessibility across two highly diverged species, our results indicate that a majority of genes involved in cell-type development are unique to each lineage. This finding contrasts with greater conservation of *cis*-regulatory dynamics, as a greater proportion of TF motifs exhibited consistent spatiotemporal progressions. Lastly, we anticipate

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

that the dissemination of ACRs by cell-type presents an ideal framework for future crop improvement efforts via genome editing and synthetic biology approaches.

ACKNOWLEDGEMENTS

We would like to acknowledge Dr. Zefu Lu for training in nuclei extraction, Julie Nelson from the University of Georgia flow cytometry core for helping sort nuclei, Noah Workman and Julia Portocarrero from the Georgia Genomics and Bioinformatics Core for library preparation, the Georgia Advanced Computing Resource Center for providing access to computing resources, and Tyler Earp for assistance with data management. This study was funded by support from the National Science Foundation (IOS-1856627 and IOS-1802848), the Pew Charitable Trusts and the UGA Office of Research to RJS, the National Science Foundation (IOS-1916804) to AG, and an NSF Postdoctoral Fellowship in Biology (DBI-1905869) to APM. AG and RJS are also supported by an NSF Collaborative award to support this research (IOS-2026554 and IOS-2026561). RJS is a Pew Scholar in the Biomedical Sciences, supported by The Pew Charitable Trusts.

AUTHOR CONTRIBUTIONS

APM and RJS designed the research. APM and ZC performed the experiments. APM, ZC, AG and RSJ analyzed the data. APM and RJS wrote the manuscript.

COMPETING INTERESTS

RJS is a co-founder of REquest Genomics, LLC, a company that provides epigenomic services. APM, ZC and AG declare no competing interests.

SUPPLEMENTARY INFORMATION

Supplementary information is available for this paper.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

CORRESPONDENCE

Correspondence and requests for materials should be addressed to Robert J. Schmitz (schmitz@uga.edu) and Alexandre P. Marand (marand@uga.edu).

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

METHODS

Plant material and 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%. 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 (**Supplementary Table 1**).

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.

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

placed in sterile water and used for nuclei isolation. 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.

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 and ~30 axillary buds). One scATAC-seq library (B73 seedling) was derived from flash frozen tissue (liquid nitrogen, followed by 7-day -80°C storage), while the remaining nine libraries were constructed with freshly harvested tissue (**Supplementary Table 1**).

To isolate individual plant nuclei, fresh or flash frozen tissue from multiple biological samples were placed on petri dishes and vigorous chopped with a 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 manufacture's instruction (10X Genomics, catalog: 1000176). Libraries were sequenced with Illumina NovaSeq 6000 in dual-index mode with six and 16 cycles for i7 and i5 index, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

386

387 ***In situ* hybridizations**

388 3-4mm tassel and ear primordia and young seedlings from the maize B73 inbred line were
389 dissected and fixed in a cold paraformaldehyde acetic acid solution (4% PFA) for 48 hours.
390 Following dehydration through a graded ethanol series and clearing of the tissue with a Histo-
391 clear II solution (Electron Microscopy Sciences), samples were embedded using Paraplast Plus
392 tissue embedding media (McCormick Scientific). 8mm sections were hybridized at 56°C with
393 antisense probes labelled with digoxigenin (DIG RNA labeling mix, Roche), and detected using
394 NBT/BCIP (Roche). Probes were synthesized by *in vitro* transcription (T7 RNA polymerase,
395 Promega) of PCR products obtained from embryo cDNA or from digested full-length cDNA clones.
396 The vectors and primers used for probe design are listed in **Supplementary Table 6**.

397

398 **Raw reads processing**

399 The following data processing was performed using each tissue and/or replicate independently
400 unless noted otherwise. Raw BCL files were demultiplexed and convert into fastq format using
401 the default settings of the 10X Genomics tool *cellranger-atac makefastq* (v1.2.0). Partial raw read
402 processing (adapter/quality trimming, mapping and barcode attachment/correction) was carried
403 out with *cellranger-atac count* (v1.2.0) using AGPv4 of the maize B73 reference genome[43].
404 Properly paired, uniquely mapped reads with mapping quality greater than 10 were retained using
405 *samtools view* (v1.6; -f 3 -q 10) and by filtering reads with XA tags[44]. Duplicate fragments were
406 collapsed on a per-nucleus basis using *picardtools* (http://broadinstitute.github.io/picard)
407 *MarkDuplicates* (v2.16; BARCODE_TAG=CB REMOVE_DUPLICATES=TRUE). Reads mapping
408 to mitochondrial and chloroplast genomes were counted for each barcode, then excluded from
409 downstream analysis. We removed reads representing potential artifacts by excluding alignments
410 coincident with a blacklist of regions composed of low-complexity and homopolymeric sequences
411 (*RepeatMasker* v4.07)[45], nuclear sequences with homology (greater than 80% identity and

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

coverage) to mitochondrial and chloroplast genomes (*BLAST*+ v2.7.1)[46], 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[10]. 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.

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[47]. 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

438 data. First, nuclei below two standard deviations from the mean fraction of reads mapping to within
439 2-kb of TSSs were removed on a per-library basis. Then, we estimated TSS enrichment scores
440 by calculating the average per-bp coverage of 2-kb windows surrounding TSSs, scaling by the
441 average per-bp coverage of the first and last 100-bp in the window (background estimate; average
442 of 1-100-bp and 1901-2000-bp), and smoothing the scaled signal with rolling-means (R package;
443 *Zoo*). Per barcode TSS enrichment scores were taken as the maximum signal within 250-bp of
444 the TSS. Lastly, for each library, we removed any barcode with a proportion of reads mapping to
445 chloroplast and mitochondrial genomes greater than two standard deviations from the mean of
446 the library.
447
448 **Detection of multiplet droplets**
449 To estimate the empirical proportion of doublets present in our data, we demultiplexed the two-
450 genotype (B73 and Mo17) pooled seedling scATAC-seq sample and assessed the proportion of
451 barcodes reflecting a mixtures of reads derived from both genotypes. Specifically, B73 and Mo17
452 whole genome short read resequencing data were acquired from PRJNA338953. Paired-end
453 reads were quality and adapter trimmed with *fastp* (v0.19.5)[48] and aligned to the B73 v4 maize
454 reference genome[43] using *BWA mem*[49] with non-default settings (-MT 1). Duplicate reads were
455 removed using *samtools rmdup*[44] (v1.6). The genomic coordinates of short nucleotide variants
456 (SNVs; single nucleotide polymorphisms [SNPs] and small insertions/deletions [INDELs]) for both
457 genotypes were identified using *freebayes*[50] (v1.0.0) with non-default settings (--min-repeat-
458 entropy 1 --min-alternate-fraction 0.05). Only biallelic SNPs – requiring at least 5 reads per
459 genotype where B73 and Mo17 were homozygous for reference and alternate nucleotides,
460 respectively – were retained. Genotypes were called by modeling allele counts as a binomial
461 distribution with a term accounting for the sequencing error rate, E_t (determined empirically as
462 the fraction of SNPs failing to match either allele), estimating posterior probabilities via Bayes

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

theorem, and assigning the genotype (or mixture of genotypes) with the greatest probability (Eq. 1-7). 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) (Eq. 1-3):

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

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

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

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 (Eq. 4-6):

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

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

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

***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[1]. 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 Sematic Indexing, LSI). Each row was centered and standardized, capping the values at $\pm$ 1.5. Crude clusters were visually identified (k=[6..12]) 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

integrations, and running *MACS2*[51] 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 (**Extended Data Fig. 1b**) 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[25].

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 less than 50 accessible ACRs. Inspired by recent developments in modeling single-cell RNA-seq data[52], 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

539 5,000 representative ACRs selected by kernel density sampling of ACR usage (fraction nuclei
540 that are accessible at a given ACR) were used to model the parameters of each ACR, using ACR
541 usage as a covariate in a generalized linear model. Specifically, the expected accessibility of an
542 ACR, y_i, can be estimated with a generalized linear model containing a binomial error distribution
543 and logit-link function, and an overdispersion term with a quasibinomial probability density
544 function (Eq. 8).
545
546 8. $\mathbb{E}(\mathrm{y}_i) \sim \beta_0 + \beta_1 log10(t)$
547
548 Where t is a vector of the sums of accessible ACRs across cell j (Eq. 9):
549
550 9. $t = \sum_i \mathrm{y}_{ij}$
551
552 To prevent over-fitting and ensure robust estimates in light of sampling noise, we learned the
553 global regularized model parameters, including overdispersion, using the representative ACRs by
554 fitting each parameter against the $\log_{10}$ fraction of accessible nuclei via kernel regression,
555 resulting in smoothed parameter estimates across the spectrum of ACR accessibility penetrance
556 present in these data. The learned global regularized model parameters were then used to
557 constrain fitted values across all ACRs for each nucleus with a simple affine transformation. To
558 account for technical variation among nuclei (variation in barcode $\log_{10}$ transformed read-depth,
559 in particular) we calculated Pearson residuals for each ACR, scaling the residuals by the
560 regularized dispersion estimate and centering values via mean subtraction, representing
561 variance-stabilized and read-depth normalized values of accessibility for a nucleus at a given
562 ACR. We note that this method is amenable to calculating residuals that account for additional

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

563 sources of technical variation, including categorical and numeric covariates, that may obscure
564 biological signal, such as batch effects, proportion of mitochondrial reads, etc.

565 The dimensionality of the Pearson residual matrix was reduced using singular value
566 decomposition (SVD) implemented by the R package *irlba*[53], retaining the first 25 left singular
567 vectors scaled by singular values (hereafter referred to as nuclei embeddings), analogous to
568 principal components (PCs) on an uncentered matrix. Nuclei embeddings were then standardized
569 across components and filtered to remove components correlated with barcode read depth
570 (Spearman's rho > 0.7). We further reduced the dimensionality of the nuclei embedding with
571 Uniform Manifold Approximation Projection (UMAP) via the R implementation of *umap-learn*
572 (min_dist = 0.1, k=50, metric="euclidean"). Nuclei were clustered with the *Seurat* v3[54] framework
573 and Louvain clustering on a k=50 nearest neighborhood graph at a resolution of 0.02 with 100
574 iterations and 100 random starts. Clusters with aggregated read depths less than 1.5M were
575 removed. To filter outliers in the UMAP embedding, we estimated the mean distance for each
576 nucleus with its *k* (*k*=50) nearest neighbors and removed nuclei greater than 3 standard deviations
577 from the mean.

578 We observed fine-scale heterogeneity within major clusters, thus we repeated our
579 clustering pipeline for each major cluster independently by partitioning the SVD embedding into
580 the top 20 components, L2 normalizing nuclei embeddings across components, and projecting
581 the L2-normalized embeddings into the UMAP space. Subclusters of nuclei were identified by
582 Louvain clustering on the L2 normalized SVD embedding (resolution set manually, range = 0.6 –
583 1.0) with 20 nearest neighbors, filtering outlier nuclei more than 2 standard deviations from the
584 mean distance of 25 nearest neighbors within each cluster.

585 For analysis of chromatin accessibility across clusters, we assembled a matrix of clusters
586 by ACRs by aggregating the number of single-base resolution Tn5 integration sites from nuclei
587 within the same cluster for each ACR, analogous to normalizing by the proportion of reads in
588 peaks for each cluster. To account for differences in read depth and other technical factors, the

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

raw counts were transformed with *edgeR's* "cpm" (log=T, prior.count=5) as previously described[55]. 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[24]. A similar framework was applied to populations of single cells, which showed that co-accessible ACRs are typically more conserved and functionally associated[26]. To identify potentially functional co-accessible ACRs, we applied a recently developed method, *Cicero*[23], 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 data sets derived from maize seedling and pistillate inflorescence primordia[10].

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

615

616 **Estimation of gene accessibility scores**

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

630 Relative gene accessibility scores exhibited a bimodal distribution with relative gene
631 accessibility values near zero resembling low or non-expressed genes. We applied a gaussian
632 mixture-model (two distributions) based scaling step per cluster to reduce noise introduced by
633 genes with low gene accessibility. Briefly, the average gene accessibility across nuclei was fit to
634 a two distribution gaussian mixture model in each cluster using the R package *mclust*. We
635 estimated cluster-specific scaling parameters determined as the 5% quantile of non-zero gene
636 accessibility values of genes from the gaussian distribution with the larger mean, for each cluster.
637 This parameter was then used to scale gene accessibility scores for all genes in each nucleus
638 within the cluster. Scaled gene accessibility scores were rounded to the nearest integer and
639 normalized across all nuclei and clusters using nucleus-specific size factors estimated as the total
640 gene accessibility of a nucleus divided by the exponential of the mean of log-transformed gene

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

641 accessibility sums across nuclei. To aid visualization, we smoothed normalized gene accessibility
642 scores by estimating a diffusion nearest neighbor graph (k=15) using the SVD embedding with 3
643 steps similar to the *run_MAGIC* function in R package *snapATAC*[56]. Downstream analyzes based
644 on binarized gene accessibly were conducted by simply converting normalized (non-smoothed)
645 accessibility scores to 1 for all positive values.
646
647 **Cell-type annotation**
648 To identify and annotate cell-types for each barcode, we identified marker genes known to localize
649 to discrete cell-types or domains expected in the sampled tissues/organs based on extensive
650 review of the literature (**Supplementary Table 2**). To enable gene accessibility comparisons
651 among clusters, we generated three pseudo-replicates for each cluster by resampling nuclei
652 within the cluster such that all cluster pseudo-replicates contained the mean number of nuclei
653 across clusters (number of nuclei per pseudo-replicate = 552) without replacement when possible.
654 To identify genes with increased accessibly relative to other clusters, we constructed a reference
655 panel with three pseudo-replicates by uniformly sampling nuclei without replacement from each
656 organ (number of nuclei per organ = 92), with a total of 552 nuclei per reference panel pseudo-
657 replicate. Read counts per gene were summed across nuclei within each pseudo-replicate. Using
658 the *DESeq2* R package, we identified genes with significantly different (FDR < 0.01) accessibility
659 profiles between each cluster and the reference panel.
660 The list of significantly differentially accessible genes was filtered to retain the genes on
661 our list of cell-type specific markers. We initially ranked the top three marker genes in each cluster
662 by their test statistics. To account for clusters containing small proportions of contaminating nuclei
663 of a different cell-type, we adjusted the test statistics using a previously described method[1],
664 effectively scaling marker activity scores by the proportion of nuclei in the cluster that were derived
665 from an organ in which the marker gene i is an expected cell-type. Clusters where the top three
666 markers corresponded to the same cell-type were annotated with the consensus cell-type.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 deviated 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*[57], scaling gene accessibility scores in each nucleus with counts per million. Next, we estimated cell-type

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

693 enrichment scores for each nucleus by calculating the mean accessibility scores of markers for a
694 given cell-type, subtracting the mean background signal defined as 1,000 sets of averaged
695 randomly sampled genes (each set had the same number of genes as the number of markers),
696 divided by the standard deviation of the background signal. Enrichment scores for each nucleus
697 were transformed into a probability distribution by dividing by the sum of cell-type enrichment
698 scores. For each nucleus, we compared the top two most likely cell-types, retaining nuclei where
699 the top predicted cell-type had a two-fold greater probability than the next most likely assignment.
700 We used these high-confidence cells to train a regularized logistic multinominal classifier with the
701 R package, *glmnet*. Cell-type classifications with less than 10 nuclei in the training set were
702 excluded. We used a LASSO L1 penalty to regularize the logistic classifier, modeling the training
703 set of nuclei as observations and TMM gene accessibility scores as variables. We balanced
704 observations by weighting by the inverse frequency of cell-types in the training set. The model
705 was trained with 10-folds and evaluated by testing a 20% hold-out set of nuclei. The predicted
706 cell-type for each nucleus in the atlas was taken as the cell-type with the greatest probability if
707 the probability ratio between the best and next best assignment was greater than five-fold,
708 otherwise labeled as 'unknown'. Using these per-cell assignments, we defined subclusters as the
709 majority cell-type if greater than 50% of nuclei in the cluster were in agreement, labelling clusters
710 with two or more majority cell-types as 'mixed' and all other clusters as 'unknown'. All cell-type
711 labels from these three automated approaches were manually reviewed by careful evaluation with
712 UMAP gene accessibility score embeddings and cluster aggregated coverages for all marker
713 genes and refined *ad hoc*.
714
715 **STARR-seq analysis**
716 Single bp-resolution enhancer activities were available from a previous study[10]. Enhancer activity
717 (defined as the $\log_2$ ratio between RNA and DNA input fragments scaled per million) for each ACR
718 was taken as the maximum over the entire ACR. A control set of regions was generated to match

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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. Specifically, for each cluster, we compared binarized ACR accessibility scores to a reference panel of uniformly sampled nuclei from each organ (111 nuclei from each organ) where the total number of reference nuclei was set to the average number of nuclei per cluster (n=666). We then fit two generalized linear logistic regression models (Eq. 10-11), with and without a term for membership to the cluster of interest.

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

743 considered as cluster-specific. Analysis of differential gene accessibility was performed as
744 described in the section titled “Cell-type annotation”.
745
746 **GO gene set enrichment analysis**
747 Gene set enrichment using GO biological process terms was performed using the R package
748 *fgsea.* For each cluster, test statistics were multiplied by the sign of the $\log_2$ fold-change value
749 versus the reference panel. GO terms with gene sets less than 10 and greater than 600 were
750 excluded from the analysis. GO terms were considered significantly enriched at FDR < 0.05
751 following 10,000 permutations.
752
753 **Motif analysis**
754 Motif occurrences were identified genome-wide with *fimo* from the MEME suite toolset[58] using
755 position weight matrices (PWM) based on DAP-seq data in *A. thaliana* and *Zea mays*[59,60]. To
756 identity TF motifs associated with cell-type specific ACRs, we ranked the top 2,000 ACRs in each
757 cell-type by Z-scores derived from CPM normalized accessibility values (see section above
758 “Nuclei clustering”). As a reference for comparison, we identified 2,000 “constitutive” ACRs that
759 varied the least and were broadly accessible across clusters. The number of ACRs containing a
760 specific motif was compared to the frequency of constitutive ACRs harboring the same motif using
761 a binomial test for each cell-type and motif. To control for multiple testing, we used the Benjamini-
762 Hochberg method to estimate the FDR, considering tests with FDR < 0.05 as significantly different
763 between the focal cell-type and constitutively accessible regions. Maize homologs of *A. thaliana*
764 TFs were identified using protein fasta alignments from BLAST+ v2.10.0 with an E-value cut-off
765 of 1e-5. Only fasta sequences classified as transcription factors from either species were
766 considered during alignment. To narrow the list of putative orthologs based on functional similarity
767 to *A. thaliana* TFs, we filtered alignments with less than 30% identity, removed maize TFs
768 classified as belonging to a different family, and selected the homolog with the greatest Pearson

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

769 correlation coefficient (PCC) with respect to the motif deviation score. Motif deviation scores of
770 specific TF motifs among nuclei were estimated using *chromVAR*[61] with the non-redundant core
771 plant PWM database from JASPAR2018. The input matrix for *chromVar* was filtered to retain a
772 minimum of 50 accessible nuclei per ACR and barcodes with at least 50 accessible ACRs. We
773 visualized differences in global motif usage per nucleus by projecting deviation scores onto the
774 UMAP embeddings. To determine if patterns of TF motif accessibility from individual nuclei could
775 be used to predict cell-type annotations, we constructed a neural network for multinomial
776 classification using the R package, *caret*[62] (method="multinom") using 80% of nuclei to train, 10-
777 fold cross-validation, averaging error terms across 10 iterations. The nuclei in the 20% withheld
778 group were used to test the model. Sensitivity, specificity and accuracy of the model was
779 evaluated using the function confusionMatrix from *caret*.
780
781 **Analysis of cell-type specific selection signatures**
782 Multi-locus allele-frequency differentiation signals between chronologically sampled elite maize
783 inbred lines were mapped onto ACRs[19], where the selection score for an ACR was taken as the
784 maximum XP-CLR value within the 500-bp ACR interval. To identity cell-types associated with
785 increased signatures of selection, the top 2,000 ACRs defined by standardized quantile-scaled
786 CPM chromatin accessibility (Z-scores, see above "Nuclei clustering") were identified for each
787 cell-type. The mean XP-CLR scores per-cell type were standardized by the mean and standard
788 deviation of randomly sampled ACRs (n=2,000) without replacement across 1,000 permutations,
789 where each permutation estimates the mean XP-CLR scores of a random subset of 2,000 ACRs
790 from the total list of 165,913 possible ACRs. Enrichment *Z*-scores were converted into *P*-values
791 using the R function *pnorm* (log.p=T, lower.tail=F) and used to estimate FDR via the Benjamini-
792 Hochberg method with the R function *p.adjust* (method="fdr").
793
794 **Analysis of co-accessible ACRs**

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To enable comparison with previously identified Hi-C and HiChIP loops[10], 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 identity 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").

Pseudotime analysis

Pseudotime trajectories were constructed similar to previous methods[63]. 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

821 analysis (ACRs, motifs, and TF activity) across pseudotime, nuclei were sorted by ascending
822 pseudotime. The ACR x nucleus matrix was filtered to retain differentially accessible ACRs (see
823 section “Analysis of differential accessibility across pseudotime” below) with at least one nucleus
824 defined as accessible. For each ACR, we fit a generalized additive model with the binary
825 accessibility scores as the response and a smoothed pseudotime component as the dependent
826 variable [s(pseudotime, bs=”cs”)] with a binomial error term and a logit-link function with *gam* from
827 the *mgcv* R package. Predicted accessibility scores across pseudotime were generated from 500
828 equally spaced interpolated points covering the range of pseudotime values. Finally, predicted
829 accessibility scores were mean-centered, standardized and constrained to the range ± 1 for each
830 ACR. Model specification for motif deviations and TF gene accessibility analysis was similar to
831 ACR pseudotime analysis with the exception of a Gaussian error distribution, and TF gene
832 accessibility was normalized by the row maximum rather than rescaling on a ± 1 distribution.
833
834 **Analysis of differential accessibility across pseudotime**
835 To identity differentially accessible ACRs across pseudotime, we fit normalized accessibility
836 residuals (Pearson residuals from a generalized linear logistic regression model with $\log_{10}$ number
837 of accessible ACRs per barcode as the dependent variable, see section “Nuclei Clustering”
838 above) as the response and pseudotime as the dependent variable using a natural spline with six
839 degrees of freedom [ns(pseudotime, df=6)] from the R package *splines* for each trajectory. We
840 took an F-test based approach for hypothesis testing of differential accessibility across
841 pseudotime by comparing the variance explained by the splined linear model with that of the
842 residuals normalized by degrees of freedom. *P*-values from the model were used to estimate
843 Benjamini-Hochberg FDR values with the R function *p.adjust* (method=”fdr”), where a FDR
844 threshold < 0.05 denoted statistical significance for differentially accessible ACRs across
845 pseudotime. To identify genes and TF motifs with differential accessibility across pseudotime, we

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

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)[64]. Gene accessibility scores for 10,976 putative orthologs were imputed using a diffusion-based approach[56,65] 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 data set[66]. 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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

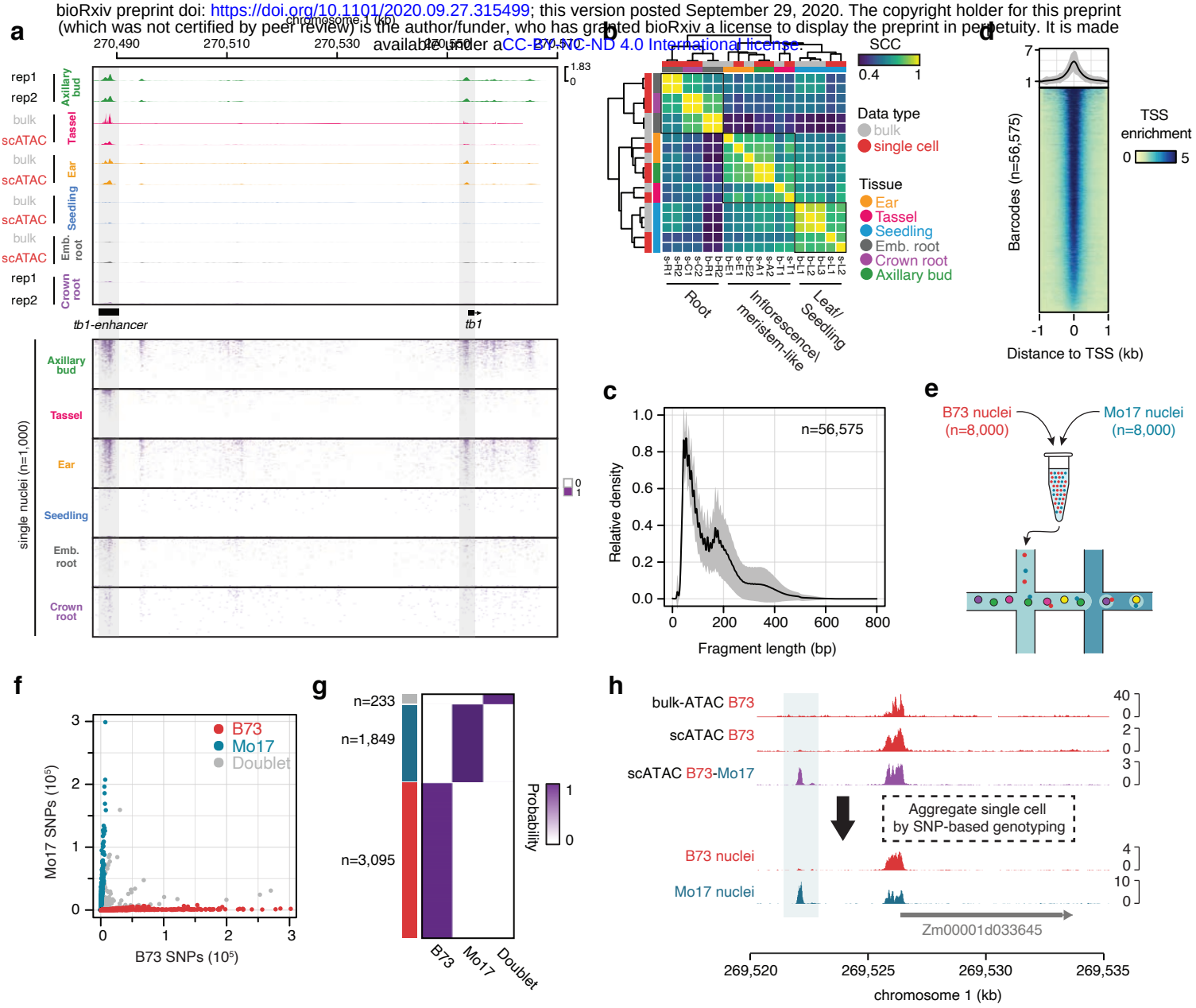
DATA AVAILABILITY

Supplementary Data are available on the online versions of this manuscript. Raw and processed data has been deposited in NCBI GEO database under accession code GSEXXXXXX. Cell-type resolved data can be viewed through our public Plant Epigenome JBrowse Genome Browser[67] (URL available upon publication).

CODE AVAILABILITY

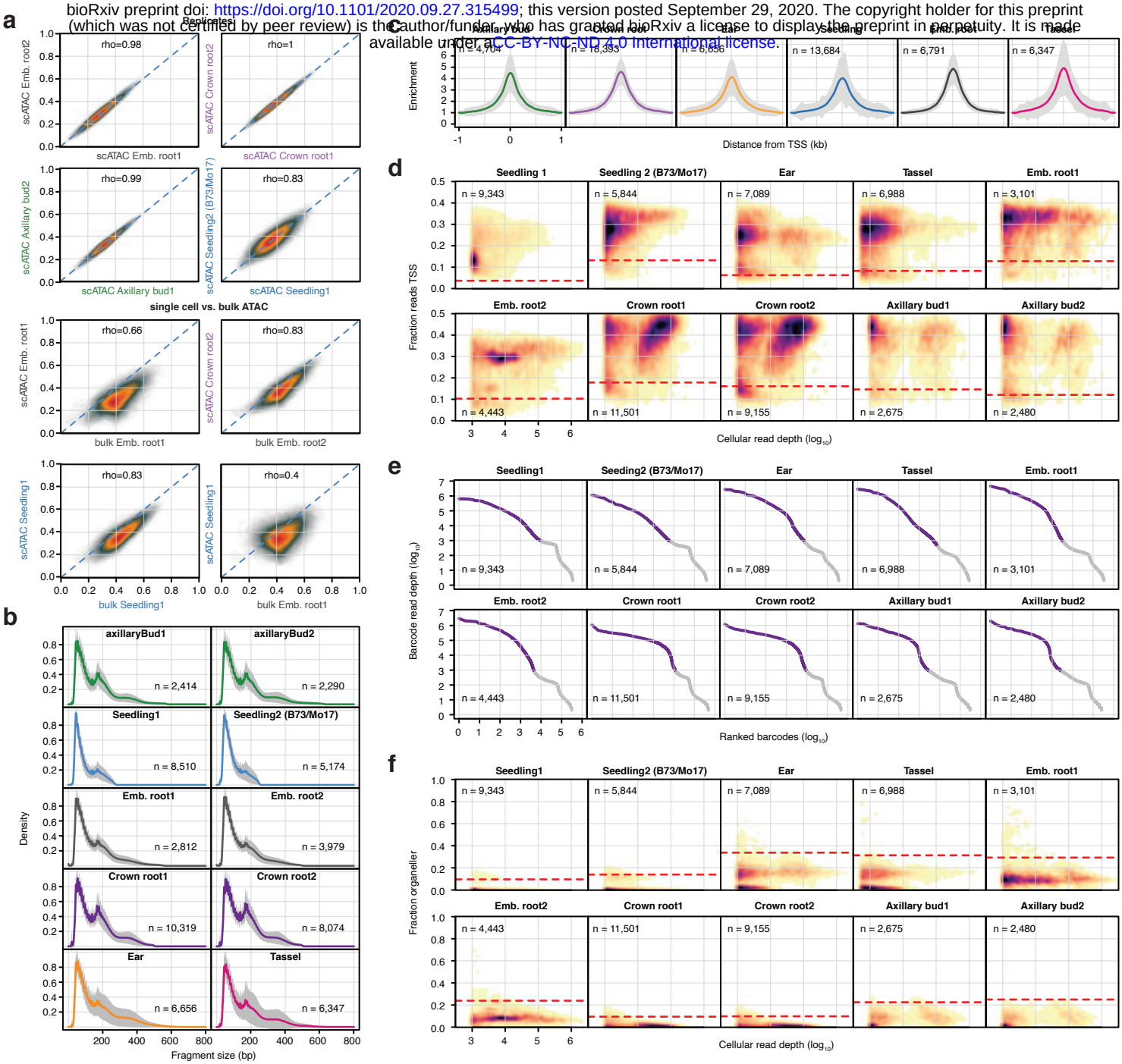
Code used throughout the analysis will be made available upon publication.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Extended Data Fig. 1: Construction of single-cell chromatin accessibility profiles from a diverse set of *Zea mays* organs. **a**, Top; Genome browser screenshot of chromatin accessibility from bulk and single-cell aggregate ATAC-seq experiments. Chromatin accessibility profiles depict the *tb1* locus and the *tb1* enhancer located approximately 67kb upstream. Bottom; binary accessibility scores from a random selection of 1,000 individual nuclei from each organ. **b**, Spearman's correlation coefficient matrix comparing bulk and pseudobulk maize ATAC-seq samples across various organs. Sample codes are short hand 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. **c**, Fragment length distributions across 56,575 cells. The solid line and grey polygon represent the average and standard deviation, respectively. **d**, Average TSS enrichment (normalized read depth adjusted by the two 10 bp windows 1kb away from TSSs) across all 56,575 cells. Grey polygon denotes the standard deviation. **e**, Genotype-mixing experimental schematic. **f**, Scatterplot of per cell B73 and Mo17 SNP counts from a mixed-genotype experiment (V1 seedlings) colored by genotype classification. **g**, Posterior probabilities of individual barcodes (rows) highlighting the occurence of cells with B73, Mo17, and mixed (doublet) genotype identities. **h**, 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 7 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 7 day old seedling sample (row 3).

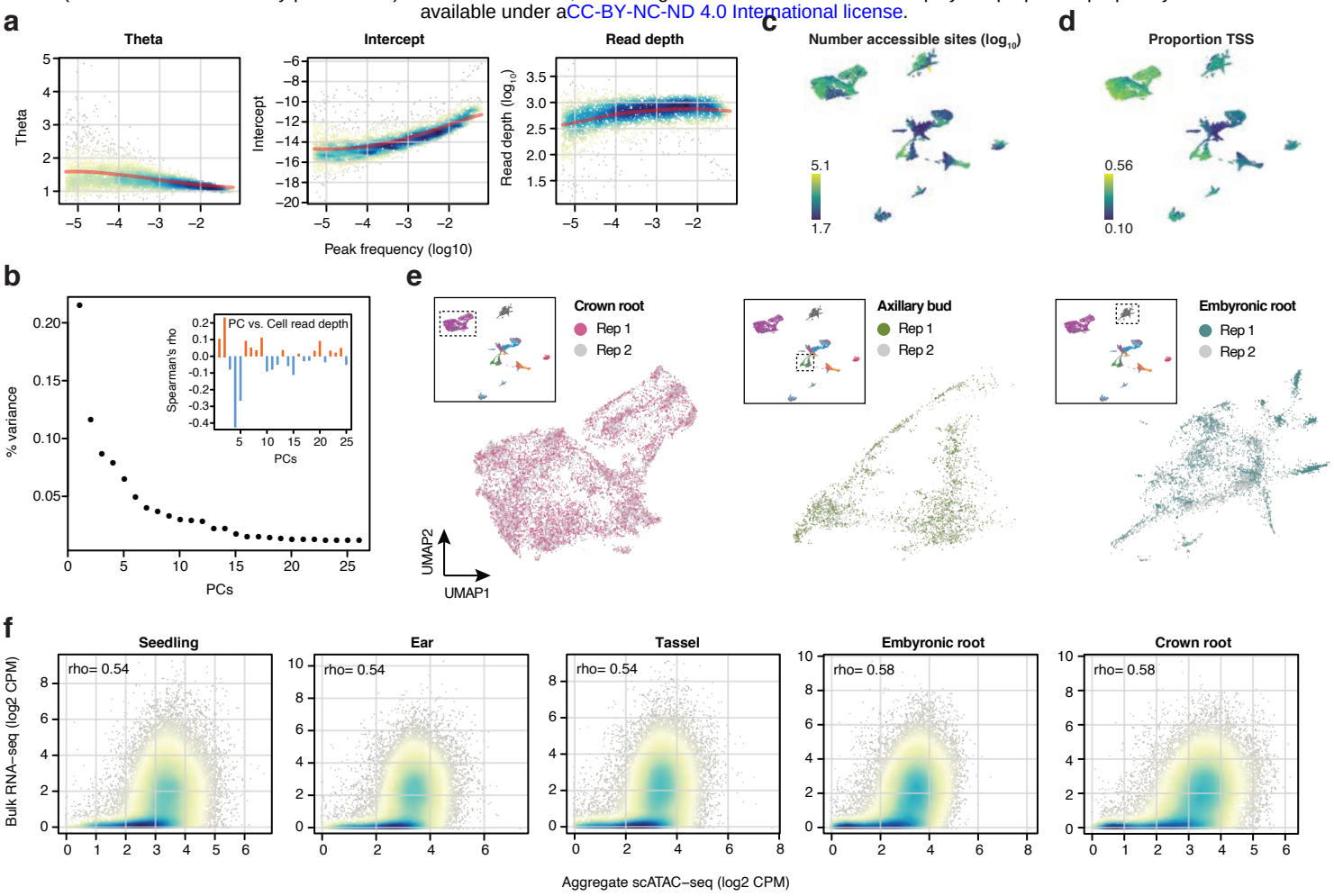
bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.



Extended Data Fig. 2: Cell calling and barcode quality control

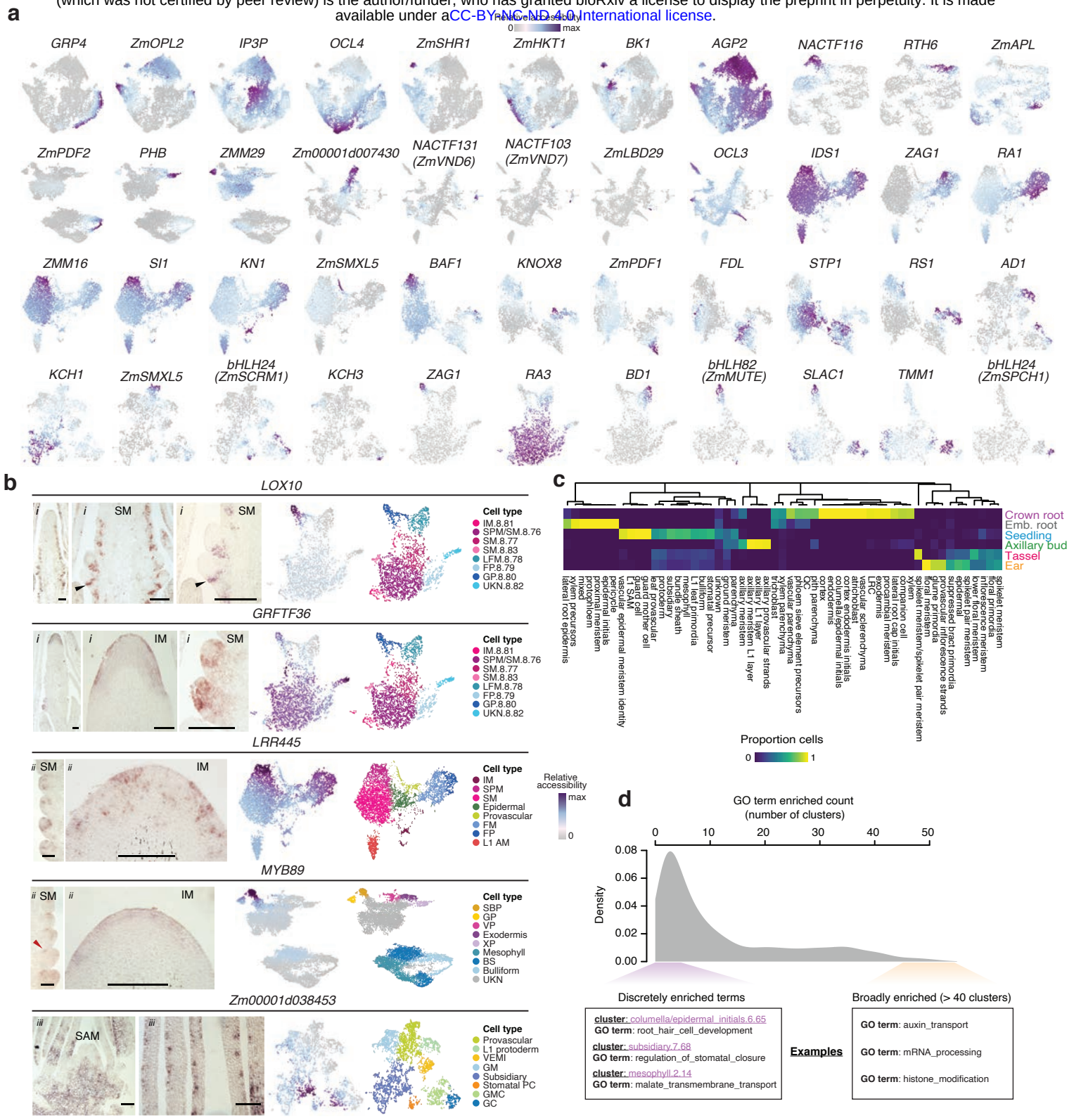
a, Comparison of normalized (0-1) read depths at the union of all peaks across bulk and single-cell samples (n=265,992) between replicated libraries, and between bulk and single-cell ATAC-seq assays. **b**, Fragment length distributions for each library. Solid lines indicate the average distribution across cells within the sample. Grey polygons represent the standard deviation across cells in the library. **c**, 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. **d**, Density scatter plots of log10 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. **e**, Knee plots illustrating log10 transformed cellular read depths of log10 ranked barcodes across libraries. **f**, Density scatter plots of log10 transformed barcode read depths (x-axis) by the fraction of Tn5 integration sites derived from organeller sequences (chloroplast and mitocondrial) relative to the total number of unique Tn5 integration sites associated with cognate barcode. Dashed red lines indicate the threshold of two standard deviations from the mean used to filter lower quality barcodes.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Extended Data Fig. 3: Clustering metrics and comparison of bulk gene accessibility and expression
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 (log10-transformed). **c**, Number of accessible sites per cell (log10). **d**, Proportion of Tn5 integrations within 2kb 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, log2 CPM) versus aggregate scATAC-seq gene accessibility scores (x-axis, log2 CPM) within an organ.

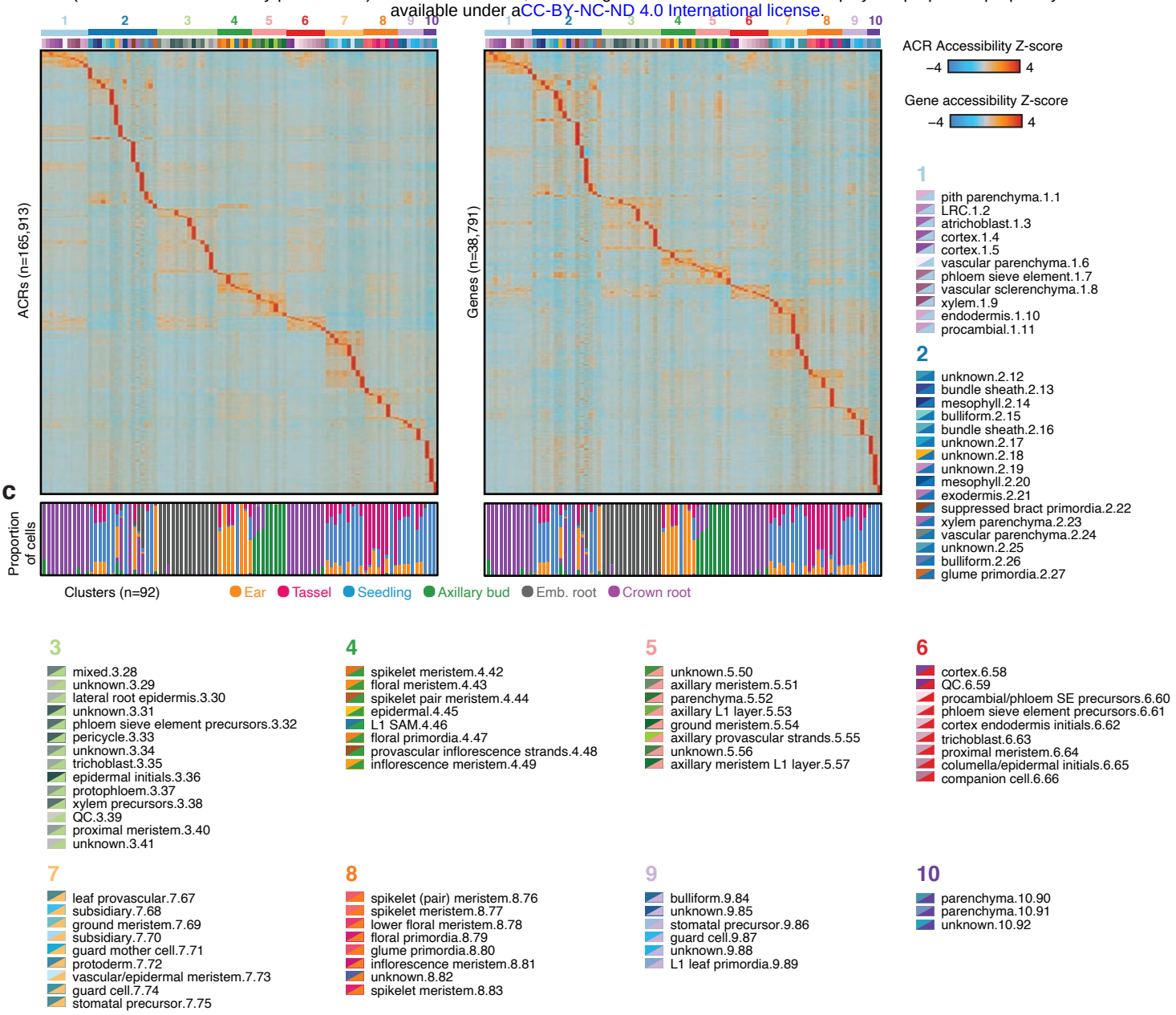
bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Extended Data Fig. 4: Cell-type annotation and GO enrichment

a, UMAP embeddings of nuclei barcodes colored by low (grey) to high (dark purple) gene activity values of cell-type specific marker genes. **b**, 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. **c,** Proportion of cells within subcluster (column) derived from one of six organs (rows). **d,** Distribution of GO term enrichment across clusters, where the x-axis indicates the number of clusters in which a GO term is significantly enriched.

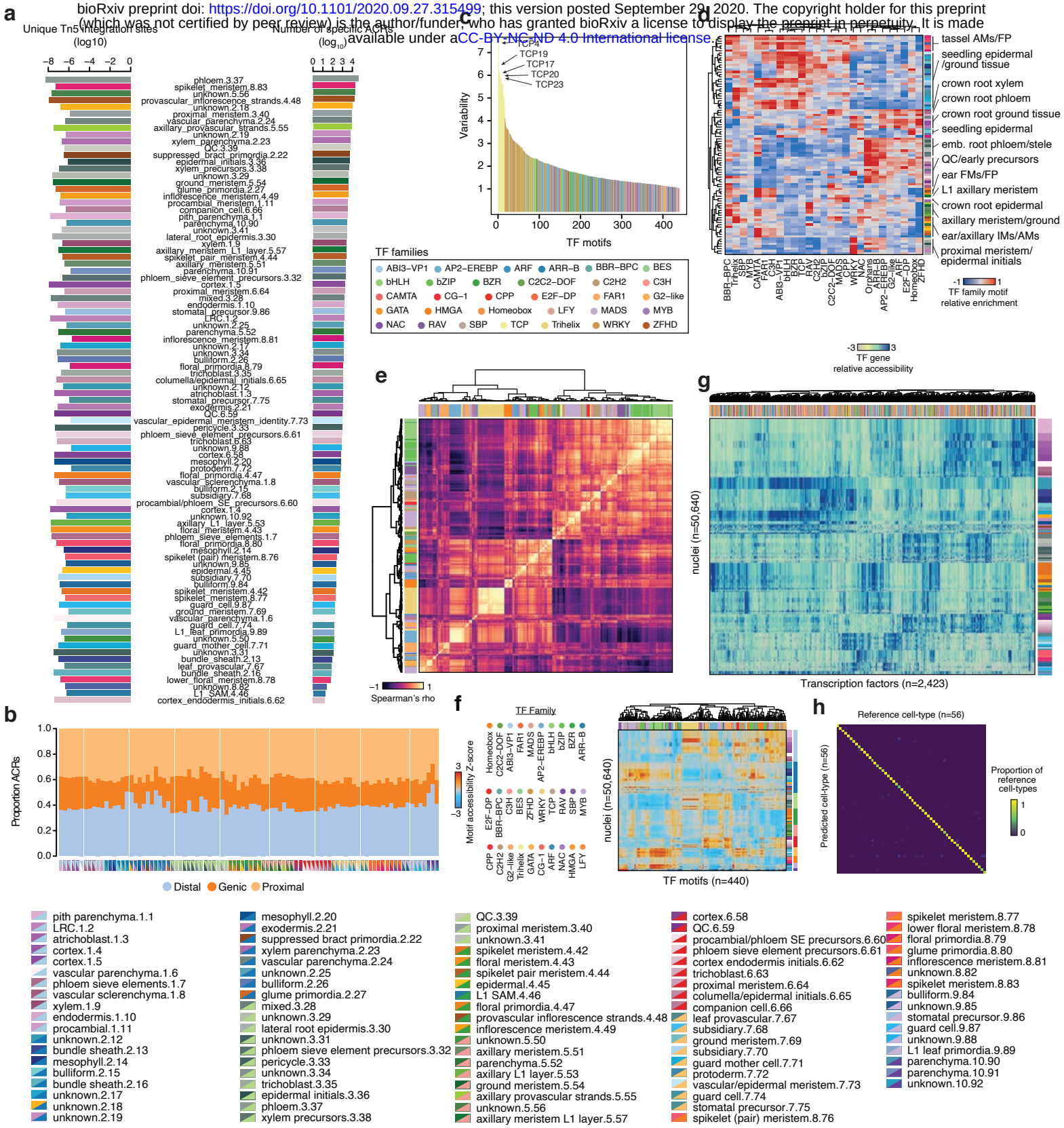
bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Extended Data Fig. 5: Chromatin accessibility variation across plant cell-types

a, Row chromatin accessibility Z-scores across cell-types. Cell-type and subcluster labels are located adjacent to the heatmap in identical order as the matrix. The first number after the cell-type label indicates the major cluster number, while the second number represents the subcluster ID. Legend colors: top left triangle, subcluster (corresponds to the sub-cluster annotation on heatmap). bottom right triangle, major cluster (corresponds to the major cluster annotation on heatmap). **b**, Row gene accessibility Z-scores across cell-types. **c**, Proportion of cells derived from one of six organs within each cell-type/subcluster.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

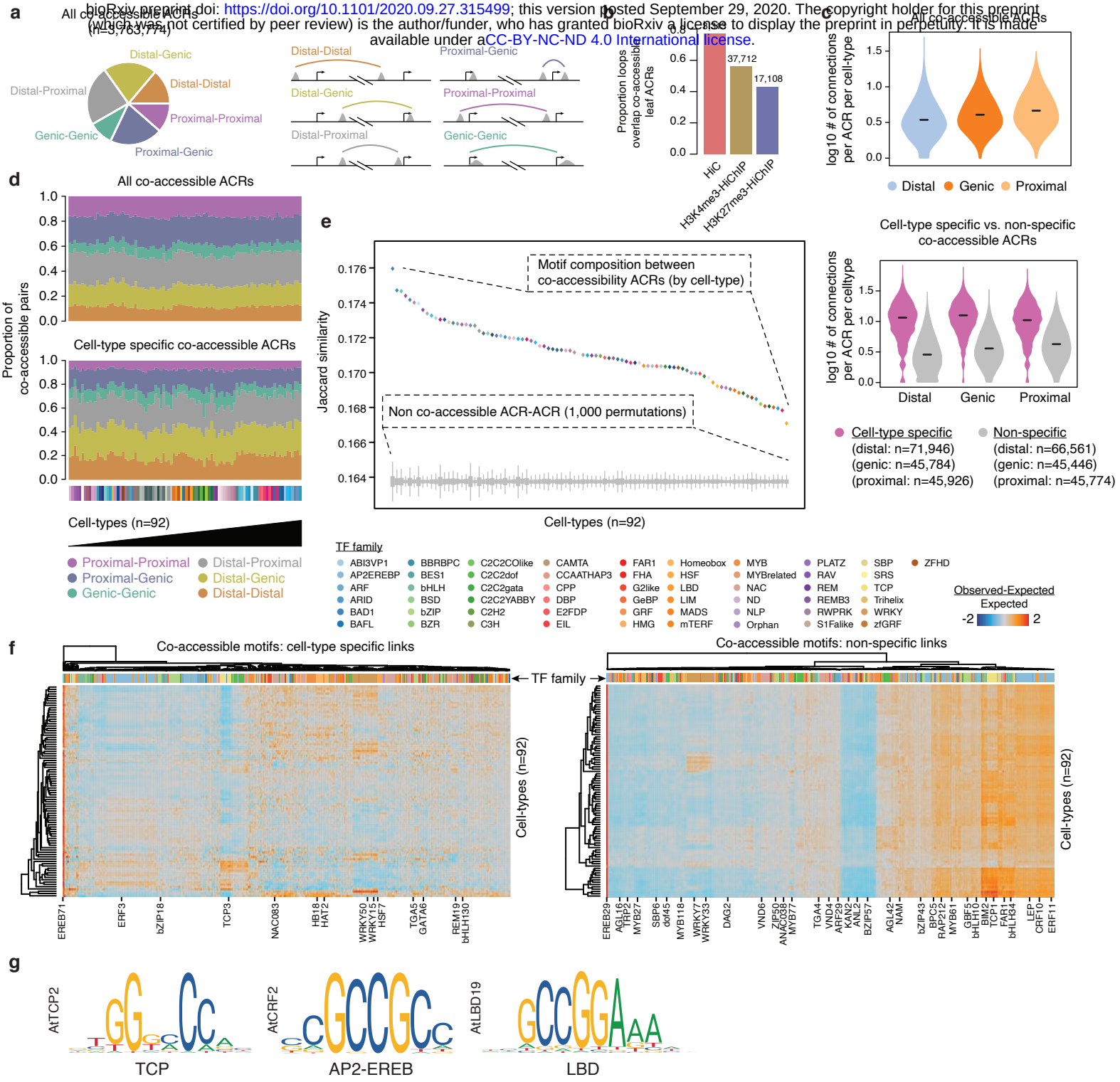


pith parenchyma.1.1
LRC.1.2
atrichoblast.1.3
cortex.1.4
cortex.1.5
vascular parenchyma.1.6
phloem sieve elements.1.7
vascular sclerenchyma.1.8
xylem.1.9
endodermis.1.10
procambial.1.11
unknown.2.12
bundle sheath.2.13
mesophyll.2.14
bulliform.2.15
bundle sheath.2.16
unknown.2.17
unknown.2.18
unknown.2.19
mesophyll.2.20
exodermis.2.21
suppressed bract primordia.2.22
xylem parenchyma.2.23
vascular parenchyma.2.24
unknown.2.25
bulliform.2.26
glume primordia.2.27
mixed.3.28
unknown.3.29
lateral root epidermis.3.30
unknown.3.31
phloem sieve element precursors.3.32
pericycle.3.33
unknown.3.34
trichoblast.3.35
epidermal initials.3.36
phloem.3.37
xylem precursors.3.38
QC.3.39
proximal meristem.3.40
unknown.3.41
spikelet meristem.4.42
floral meristem.4.43
spikelet pair meristem.4.44
epidermal.4.45
L1 SAM.4.46
floral primordia.4.47
provascular inflorescence strands.4.48
inflorescence meristem.4.49
unknown.5.50
axillary meristem.5.51
parenchyma.5.52
axillary L1 layer.5.53
ground meristem.5.54
axillary provascular strands.5.55
unknown.5.56
axillary meristem L1 layer.5.57
cortex.6.58
QC.6.59
procambial/phloem SE precursors.6.60
phloem sieve element precursors.6.61
cortex endodermis initials.6.62
trichoblast.6.63
proximal meristem.6.64
columella/epidermal initials.6.65
companion cell.6.66
leaf provascular.7.67
subsidiary.7.68
ground meristem.7.69
subsidiary.7.70
guard mother cell.7.71
protoderm.7.72
vascular/epidermal meristem.7.73
guard cell.7.74
stomatal precursor.7.75
spikelet (pair) meristem.8.76
spikelet meristem.8.77
lower floral meristem.8.78
floral primordia.8.79
glume primordia.8.80
inflorescence meristem.8.81
unknown.8.82
spikelet meristem.8.83
bulliform.9.84
unknown.9.85
stomatal precursor.9.86
guard cell.9.87
unknown.9.88
L1 leaf primordia.9.89
parenchyma.10.90
parenchyma.10.91
unknown.10.92

Extended Data Fig. 6: Transcription factor motif variation underly dynamic accessible regions

a, Comparison of cell-type specific ACR counts (by cell-type) and the number of unique Tn5 integration sites. **b,** Proportion of distal, genic and proximal ACRs per cell-type. **c,** Ranked TF motif variability across cells, colored by TF family. **d,** Mean TF family motif enrichment (average deviation scores per cell-type per TF family) across cell-types scaled to +/- 1. **e,** Spearman's correlation coefficient (rho) between TF motifs (comparison of motif deviations across all nuclei). Row and column colors represent TF motif families. **f,** TF motif deviations for 440 TF motifs (columns) per nucleus (rows). The TF family for each corresponding motif is denoted by column header colors. Cluster identification (sub-cluster/major cluster) are illustrated as row header colors on the right side of the heatmap. **g,** Relative gene accessibility scores for 2,423 maize transcription factors (columns) for 50,640 nuclei (rows). Nuclei are sorted by sub-cluster cell-type (row color labels). **h,** Proportion of annotated cell-types (columns) matching predicted cell-types from the neural network. A cell-type annotation color legend applicable to panels b, f, and g where left triangle reflects sub-cluster, right triangle indicates the major cluster.

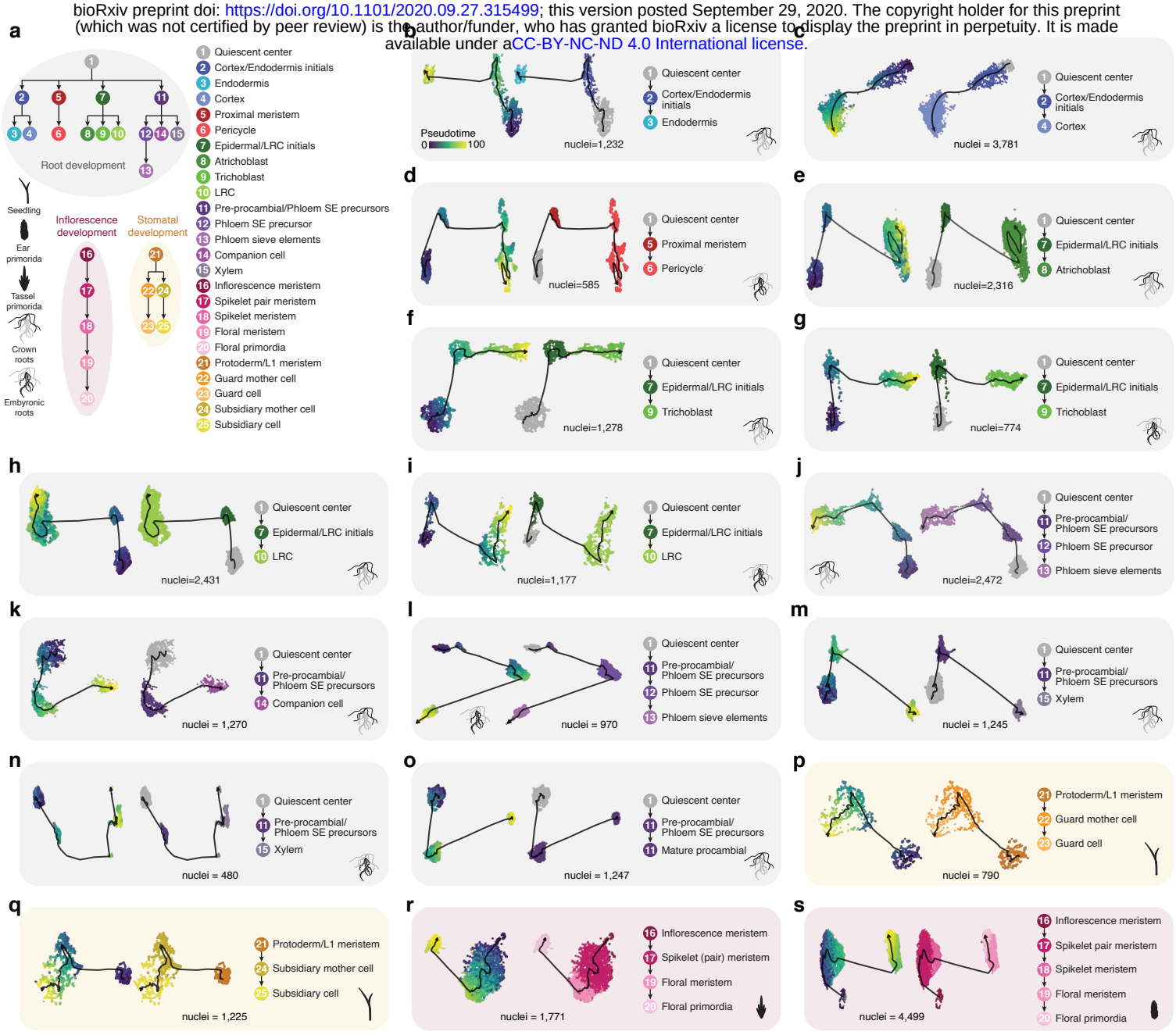
bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Extended Data Fig. 7: Co-accessible ACRs reflect in vivo chromatin interactions driven by coordinated TF activity

a, Proportions of co-accessible ACR types, illustrated by toy examples. **b**, 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 cell derived from seedlings). **c**, Top, log10 number of connections per ACR per cell-type from all co-accessible ACRs split by genomic context: distal, proximal, and genic. Bottom, log10 number of connections per ACR per cell-type from cell-type specific (purple) and non-specific (grey) co-accessible ACRs, split by genomic context. **d**, Proportion of co-accessible classifications by cell-type for all co-accessible ACRs (top) and cell-type specific co-accessible ACRs (bottom). **e**, 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 (grey boxplots). Box plots represent the interquartile range, grey lines indicate the permuted range. **f**, 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 random ACR-ACR links with the same number of co-accessible ACRs. **g,** Exemplary motifs enriched in reciprocal co-accessible ACRs for *TCP*, *AP2-EREB*, and *LBD* TF families.

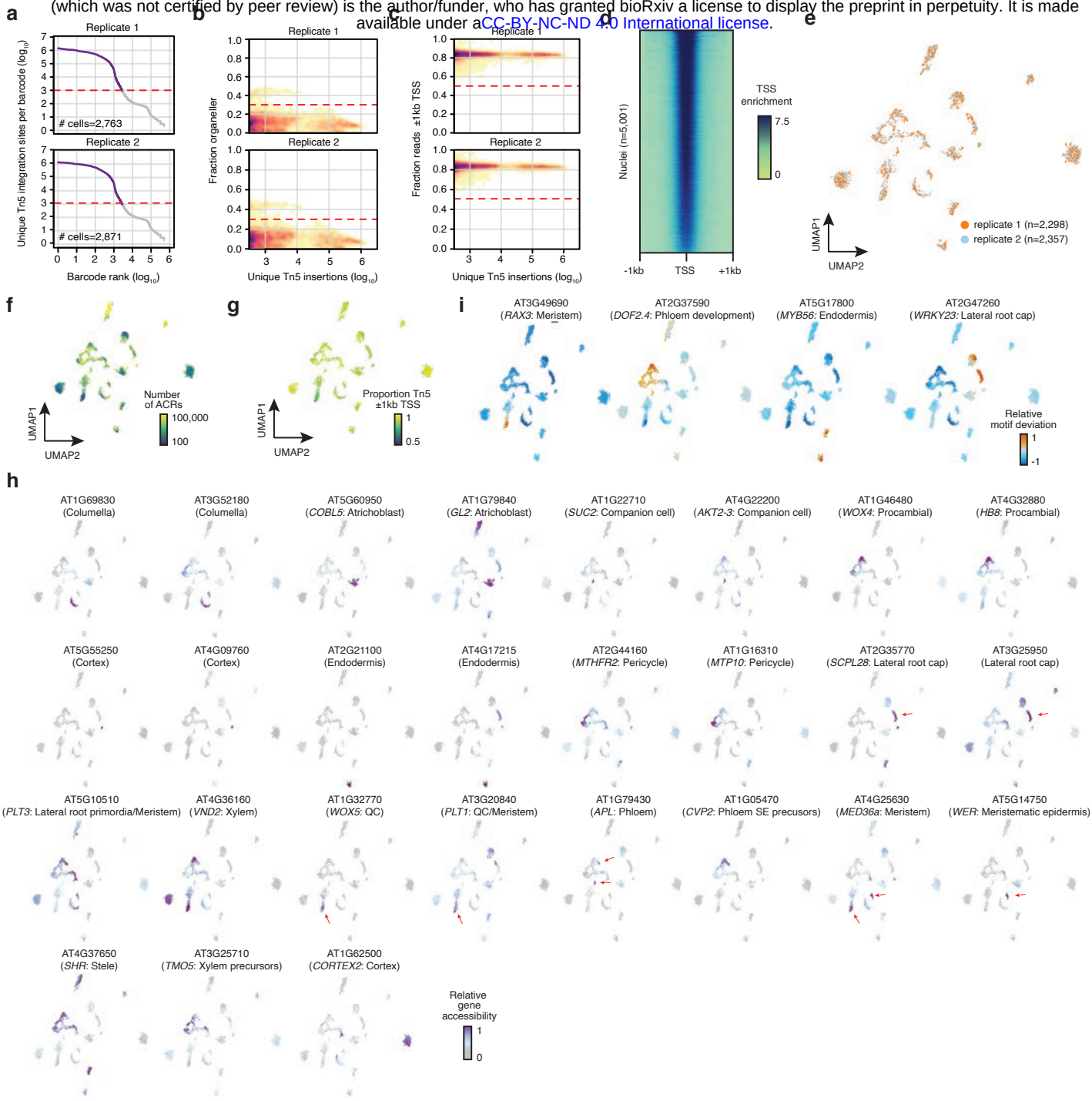
bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Extended Data Fig. 8: Pseudotime trajectory construction

a, Overview of pseudotime developmental trajectory analysis from four organs: root, seedling, tassel (staminate inflorescence), and ear (pistilate 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 pistilate inflorescence (ear).

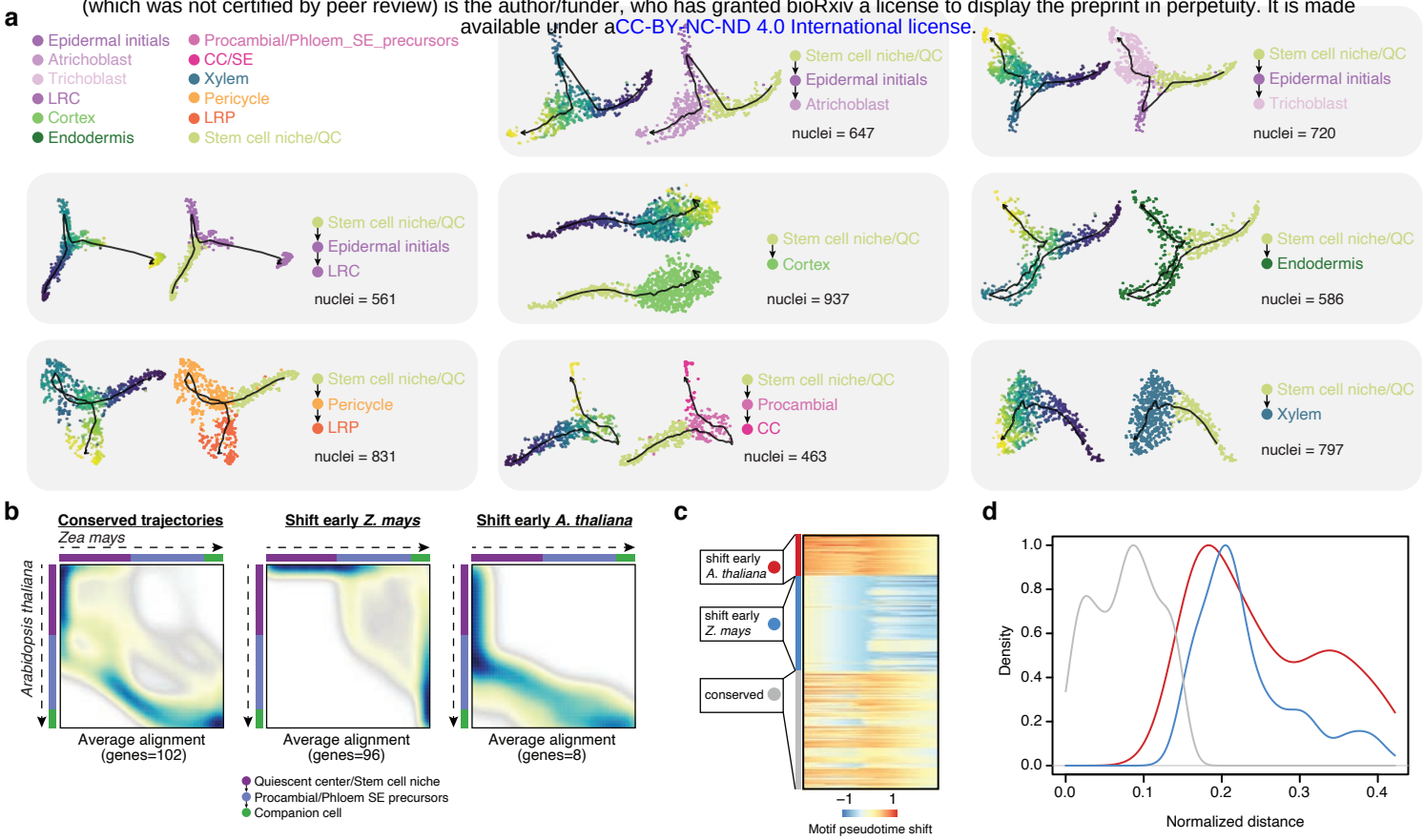
bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Extended Data Fig. 9: *Arabidopsis thaliana* root cell-type atlas

a, Knee plots for Arabidopsis thaliana root samples illustrating log10 transformed cellular read depths of log10 ranked barcodes across two biological replicates. **b,** Density scatter plots of log10 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 scatter plots of log10 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,** UMAP (Uniform manifold approximation projection) embeddings of Arabidopsis thaliana root barcodes colored by biological replicate; **f,** total number of accessible chromatin regions (ACRs); **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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Extended Data Fig. 10: Dynamic and conserved chromatin accessibility across pseudotime between *Arabidopsis thaliana* and *Zea mays*

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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

FIGURE LEGENDS

Figure 1: Atlas-scale profiling of single nuclei chromatin accessibility in *Zea mays*

a, Overview of experimental protocol. Homogenized samples release nuclei that are extracted by flow cytometry. Accessible chromatin is tagged by Tn5 and individual nuclei are distinctly barcoded in disparate droplets following microfluidic separation, allowing droplet pooling and library amplification prior to sequencing. **b,** Nuclei similarity clustering as a UMAP embedding derived from the denoised quasibinomial residuals across all ACRs for each nucleus. UMAP embedding of nuclei colored by **b,** organ identity and **c,** major cluster identity. **d,** UMAP embedding of sub-cluster assignments following a second round of clustering within each major cluster. **e,** Cell-type specific enrichment of gene accessibility for a subset of marker genes associated with six different cell-types. **f,** Sub-cluster specific chromatin accessibility profiles surrounding known marker genes for floral primordia, xylem precursors, and L1 epidermal cells. Bold, circled numbers indicate the cognate major cluster shown in panel **c**. Sub-cluster numeric identifications are present on the sides of the coverage plots. **g,** Top, gene accessibility for *ZmGRFTF36*, an inflorescence and spikelet meristem enriched TF with no previously known cell-type specificity. Bottom, RNA in situ hybridization of *ZmGRFTF36* in maize B73 staminate (tassel) primordia. FP, floral primordia; IM, inflorescence meristem; LFM, lower floral meristem; SM, spikelet meristem; SPM, spikelet pair meristem; TF, transcription factor; UKN, unknown.

Figure 2: Characterization of accessible chromatin regions

a, Distribution of enhancer activity (maximum $\log_2$[RNA/input]) for control regions (n=165,913), constitutively accessible (n=113,393) and cluster-specific ACRs (n=52,520). The dash red line indicates the overall mean. Orange lines reflect differences between the group mean and overall mean. **b,** Distribution of ACR distances to the nearest gene. Inset, distribution of ACR genomic context. **c,** Relative DNA methylation levels 2-kb flanking ACRs. **d,** Relative enrichment of

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

polymorphisms after normalizing by mapability for 5-kb regions flanking cell-type specific and constitutive ACR summits. Smoothed splines are shown as dark lines. **e,** Relative enrichment of significant GWAS polymorphisms relative to all polymorphisms for 5-kb regions flanking cell-type specific and constitutive ACR summits. **f,** Enrichment of signatures of selection (XP-CLR) in the top 2,000 ACRs for all cell-type clusters. The 20 most enriched cell-types are highlighted on the left. AM, axillary meristem; CC, companion cell; FM, floral meristem; FP, floral primordia; IM, inflorescence meristem; QC, quiescent center; SM, spikelet meristem; SPM, spikelet pair meristem. **g,** Average motif coverage for ACRs (n=165,913) and control regions (n=165,913). Shaded polygon, 95% confidence intervals. **h,** Enrichment of TF motifs in the top 2,000 ACRs ranked by Z-score for each cell-type compared to the top 2,000 most constitutive ACRs via binomial tests. FDR was estimated by the Benjamini-Hochberg method. **i,** Comparative heatmaps for matched TF gene accessibility (bottom) and motif deviation (top) Z-scores across clusters. **j,** Gene accessibility scores for five maize transcription factors (top) and their associated motif deviations (bottom).

Figure 3: Co-accessible ACRs reflect *in vivo* chromatin interactions and are established by cell-type specific TFs. a, Comparison of ear and seedling co-accessible ACRs across nine cell-types at the genetically mapped *tb1-enhancer* domestication locus (highlighted region). Link height reflects co-accessibility scores between ACRs across cells in a cluster. **b**, Average normalized Hi-C signal across 4-kb windows centered on ACRs equally distributed into three groups (each group: n=54,904) based on the number and strength of participating co-accessible links. **c**, Co-accessible ACR interaction frequency across clusters. Inset: Single cell-type co-accessible links (n=1,018,417) for different genomic contexts. **d**, Standardized (row-wise) number of connections for each ACR (rows) by cell-type (columns). Column color map reflect cell-types from the legend in Extended Data Fig. 4. Gene-proximal ACRs for a subset of marker genes are

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

948 indicated on the right. **e**, Right, chromatin accessibility and co-accessible ACR links surrounding
949 the *UB2* locus associated with ear row number and tassel branch number quantitative traits. Black
950 arrow indicates a distal ACR upstream of *UB2* present only in spikelet meristems. Co-accessible
951 links with an edge within 2-kb of *UB2* are colored pink while remaining links are grey. Link height
952 represents co-accessibility strength. Left, close-up of accessibility profiles of *UB2*. **f**, Distributions
953 of average ACR-ACR links across cell-types for ACRs that overlap (purple) and do not overlap
954 (grey) phenotype-associated genetic variants from maize GWAS. The median of each distribution
955 is shown as a white horizontal line. Violin plots present the entire range of average number of
956 connections on a log scale. Hypothesis testing was conducted within the *R* statistical framework
957 via Wilcoxon rank sum test. **g,** Distributions of average ACR-ACR links cell-types for ACRs with
958 (blue) and without (grey) enhancer activity (log_2 RNA/input greater than 0). Hypothesis testing
959 and distribution illustration was performed similarly as panel **f**.
960
961 **Figure 4: Chromatin accessibility is dynamic across pseudotime and conserved between**
962 ***Arabidopsis thaliana* and *Zea mays***
963 **a,** Overview of pseudotime trajectory analysis. Inflorescence development includes trajectory
964 construction of pistillate and staminate inflorescence, and root development encompasses
965 embryonic and crown root trajectories. **b,** UMAP embedding of companion cell developmental
966 trajectory depicting cell-types (left) and pseudotime progression (right). **c,** Relative motif
967 deviations for 440 TF motifs (left, rows), 402 TF gene accessibility scores (middle, rows), and
968 relative accessibility of 8,094 ACRs (right, rows) associated with pseudotime (columns). Four
969 motifs enriched along the trajectory gradient are shown on the left. ACRs, accessible chromatin
970 regions; TF, transcription factor. **d,** Genome browser screenshots of cell-type specific chromatin
971 accessibility profiles along the developmental trajectory for quiescent center, procambial/phloem
972 sieve element precursor, and companion cell at associated marker gene loci. **e,** UMAP
973 embedding based on whole-genome chromatin accessibility profiles of 4,655 *A. thaliana* root

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

nuclei. **f,** UMAP embedding of companion cell developmental trajectories in *A. thaliana* depicting cell-types (left) and pseudotime progression (right). Motifs, TFs, and ACRs with significant association with pseudotime are shown as heatmaps below. **g,** Per-gene pseudotime shift scores from alignments between *Z. mays* and *A. thaliana* companion cell development progressions, clustered by k-means into three groups. Distribution of gene-gene distances from the alignments are shown on the right, split by k-mean groups. **h,** Exemplary one-to-one homologs between *A. thaliana* and *Z. mays* for the three groups split by pseudotime shifts. Acronyms: CC, companion cell; LRC, lateral root cap; LRP, lateral root primordia. QC, quiescent center; SE, sieve elements.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

EXTENDED DATA FIGURE LEGENDS

Extended Data Fig. 1: Construction of single-cell chromatin accessibility profiles from a diverse set of *Zea mays* organs.

a, Top; Genome browser screenshot of chromatin accessibility from bulk and single-cell aggregate ATAC-seq experiments. Chromatin accessibility profiles depict the *tb1* locus and the *tb1* enhancer located approximately 67-kb upstream. Bottom; binary accessibility scores from a random selection of 1,000 individual nuclei from each organ. **b,** Spearman's correlation coefficient matrix comparing bulk and pseudobulk maize ATAC-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. **c,** Fragment length distributions across 56,575 cells. The solid line and grey polygon represent the average and standard deviation, respectively. **d,** Average TSS enrichment (normalized read depth adjusted by the two windows 1-kb away from TSSs) across all 56,575 nuclei. Grey polygon denotes the standard deviation. **e**, Genotype-mixing experimental schematic. **f,** Scatterplot of per cell B73 and Mo17 SNP counts from a mixed-genotype experiment (V1 seedlings) colored by genotype classification. **g,** Posterior probabilities of individual barcodes (rows) highlighting the occurrence of cells with B73, Mo17, and mixed (doublet) genotype identities. **h,** 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 7 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 7 day old seedling sample (row 3).

Extended Data Fig. 2: Cell calling and barcode quality control.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

a, Comparison of normalized (0-1) read depths at the union of all peaks across bulk and single-cell samples (n=265,992) between replicated libraries, and between bulk and single-cell ATAC-seq assays. **b,** Fragment length distributions for each library. Solid lines indicate the average distribution across cells within the sample. Grey polygons represent the standard deviation across nuclei in the library. **c,** 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. **d,** Density scatter plots 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. **e,** Knee plots illustrating $\log_{10}$ transformed cellular read depths of $\log_{10}$ ranked barcodes across libraries. **f,** Density scatter plots 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 barcode. Dashed red lines indicate the threshold of two standard deviations from the mean used to filter lower quality barcodes.

Extended Data Fig. 3: Clustering metrics and comparison of bulk gene accessibility and expression

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,

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

log_2 CPM) versus aggregate scATAC-seq gene accessibility scores (x-axis, log_2 CPM) within an organ.

Extended Data Fig. 4: Cell-type annotation and GO enrichment

a, UMAP embeddings of nuclei barcodes colored by low (grey) to high (dark purple) gene activity values of cell-type specific marker genes. **b**, 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 cryptic 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. **c,** Proportion of cells within subcluster (column) derived from one of six organs (rows). **d,** Distribution of GO term enrichment across clusters, where the x-axis indicates the number of clusters in which a GO term is significantly enriched.

Extended Data Fig. 5: Chromatin accessibility variation across plant cell-types

a, Row chromatin accessibility Z-scores across cell-types. Cell-type and subcluster labels are located adjacent to the heatmap in identical order as the matrix. The first number after the cell-type label indicates the major cluster number, while the second number represents the subcluster ID. Legend colors: top left triangle, subcluster (corresponds to the sub-cluster annotation on heatmap). bottom right triangle, major cluster (corresponds to the major cluster annotation on

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

heatmap). **b**, Row gene accessibility Z-scores across cell-types. **c**, Proportion of cells derived from one of six organs within each cell-type/subcluster.

Extended Data Fig. 6: Transcription factor motif variation underly dynamic accessible regions

a, Comparison of cell-type specific ACR counts (by cell-type) and the number of unique Tn5 integration sites. **b,** Proportion of distal, genic and proximal ACRs per cell-type. **c,** Ranked TF motif variability across cells, colored by TF family. **d,** Mean TF family motif enrichment (average deviation scores per cell-type per TF family) across cell-types scaled to +/- 1. **e,** Spearman's correlation coefficient (rho) between TF motifs (comparison of motif deviations across all nuclei). Row and column colors represent TF motif families. **f,** TF motif deviations for 440 TF motifs (columns) per nucleus (rows). The TF family for each corresponding motif is denoted by column header colors. Cluster identification (sub-cluster/major cluster) are illustrated as row header colors on the right side of the heatmap. **g,** Relative gene accessibility scores for 2,423 maize transcription factors (columns) for 50,640 nuclei (rows). Nuclei are sorted by sub-cluster cell-type (row color labels). **h,** Proportion of annotated cell-types (columns) matching predicted cell-types from the neural network. A cell-type annotation color legend applicable to panels b, f, and g where left triangle reflects sub-cluster, right triangle indicates the major cluster.

Extended Data Fig. 7: Co-accessible ACRs reflect in vivo chromatin interactions and are driven by coordinated TF activity

a, Proportions of co-accessible ACR types, illustrated by toy examples. **b**, 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 cell derived from seedlings). **c**, Top, $\log_{10}$ number of connections per ACR per cell-type from all co-accessible ACRs split by genomic context: distal,

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

proximal, and genic. Bottom, $\log_{10}$ number of connections per ACR per cell-type from cell-type specific (purple) and constitutive (grey) co-accessible ACRs, split by genomic context. **d,** Proportion of co-accessible classifications by cell-type for all co-accessible ACRs (top) and cell-type specific co-accessible ACRs (bottom). **e**, 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 (grey boxplots). Box plots represent the interquartile range, grey lines indicate the permuted range. **f**, 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 random ACR-ACR links with the same number of co-accessible ACRs. **g,** Exemplary motifs enriched in reciprocal co-accessible ACRs for *TCP, AP2-EREB,* and *LBD* TF families.

Extended Data Fig. 8: Pseudotime trajectory construction

a, Overview of pseudotime developmental trajectory analysis from four organs: root, seedling, tassel (staminate inflorescence), and ear (pistilate 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 pistilate inflorescence (ear).

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Extended Data Fig. 9: *Arabidopsis thaliana* root cell-type atlas

a, Knee plots for *A. thaliana* root samples illustrating $\log_{10}$ transformed cellular read depths of $\log_{10}$ ranked barcodes across two biological replicates. **b,** Density scatter plots 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 scatter plots 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 *A. thaliana* root barcodes (rows). **e,** UMAP (Uniform manifold approximation projection) embeddings of *A. thaliana* root barcodes colored by biological replicate. **f,** total number of accessible chromatin regions (ACRs). **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 *A. thaliana* root clusters. **i,** Relative motif deviations for transcription factors with known cell-type specificities.

Extended Data Fig. 10: Dynamic and conserved chromatin accessibility across pseudotime between *Arabidopsis thaliana* and *Zea mays*

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-

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

REFERENCES

1 Cusanovich, D. A. *et al.* A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell* **174**, 1309-1324 e1318, doi:10.1016/j.cell.2018.06.052 (2018).

2 Marand, A. P., Zhang, T., Zhu, B. & Jiang, J. Towards genome-wide prediction and characterization of enhancers in plants. *Biochim Biophys Acta Gene Regul Mech* **1860**, 131-139, doi:10.1016/j.bbagrm.2016.06.006 (2017).

3 Teale, W. D., Paponov, I. A. & Palme, K. Auxin in action: signalling, transport and the control of plant growth and development. *Nat Rev Mol Cell Biol* **7**, 847-859, doi:10.1038/nrm2020 (2006).

4 Wittkopp, P. J. & Kalay, G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet* **13**, 59-69, doi:10.1038/nrg3095 (2011).

5 Lu, Z. *et al.* The prevalence, evolution and chromatin signatures of plant regulatory elements. *Nat Plants* **5**, 1250-1259, doi:10.1038/s41477-019-0548-z (2019).

6 Andersson, R. & Sandelin, A. Determinants of enhancer and promoter activities of regulatory elements. *Nat Rev Genet* **21**, 71-87, doi:10.1038/s41576-019-0173-8 (2020).

7 Long, H. K., Prescott, S. L. & Wysocka, J. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* **167**, 1170-1187, doi:10.1016/j.cell.2016.09.018 (2016).

8 Chang, Y. M. *et al.* Characterizing regulatory and functional differentiation between maize mesophyll and bundle sheath cells by transcriptomic analysis. *Plant Physiol* **160**, 165-177, doi:10.1104/pp.112.203810 (2012).

9 Langdale, J. A., Lane, B., Freeling, M. & Nelson, T. Cell lineage analysis of maize bundle sheath and mesophyll cells. *Dev Biol* **133**, 128-139, doi:10.1016/0012-1606(89)90304-7 (1989).

10 Ricci, W. A. *et al.* Widespread long-range cis-regulatory elements in the maize genome. *Nat Plants* **5**, 1237-1249, doi:10.1038/s41477-019-0547-0 (2019).

11 Oka, R. *et al.* Genome-wide mapping of transcriptional enhancer candidates using DNA and chromatin features in maize. *Genome Biol* **18**, 137, doi:10.1186/s13059-017-1273-4 (2017).

12 Crisp, P. A. *et al.* Stable unmethylated DNA demarcates expressed genes and their cis-regulatory space in plant genomes. *Proc Natl Acad Sci U S A*, doi:10.1073/pnas.2010250117 (2020).

13 Zhao, H. *et al.* Proliferation of Regulatory DNA Elements Derived from Transposable Elements in the Maize Genome. *Plant Physiol* **176**, 2789-2803, doi:10.1104/pp.17.01467 (2018).

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

14 Clark, R. M., Wagler, T. N., Quijada, P. & Doebley, J. A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture. *Nat Genet* **38**, 594-597, doi:10.1038/ng1784 (2006).

15 Noshay, J. M. *et al.* Cis-regulatory elements within TEs can influence expression of nearby maize genes. *bioRxiv*, 2020.2005.2020.107169, doi:10.1101/2020.05.20.107169 (2020).

16 Rebeiz, M. & Tsiantis, M. Enhancer evolution and the origins of morphological novelty. *Curr Opin Genet Dev* **45**, 115-123, doi:10.1016/j.gde.2017.04.006 (2017).

17 Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160**, 554-566, doi:10.1016/j.cell.2015.01.006 (2015).

18 Wallace, J. G. *et al.* Association mapping across numerous traits reveals patterns of functional variation in maize. *PLoS Genet* **10**, e1004845, doi:10.1371/journal.pgen.1004845 (2014).

19 Wang, B. *et al.* Genome-wide selection and genetic improvement during modern maize breeding. *Nat Genet* **52**, 565-571, doi:10.1038/s41588-020-0616-3 (2020).

20 Gage, J. L., White, M. R., Edwards, J. W., Kaeppler, S. & de Leon, N. Selection Signatures Underlying Dramatic Male Inflorescence Transformation During Modern Hybrid Maize Breeding. *Genetics* **210**, 1125-1138, doi:10.1534/genetics.118.301487 (2018).

21 Verweij, W. *et al.* Functionally Similar WRKY Proteins Regulate Vacuolar Acidification in Petunia and Hair Development in Arabidopsis. *Plant Cell* **28**, 786-803, doi:10.1105/tpc.15.00608 (2016).

22 Gomez-Mena, C., de Folter, S., Costa, M. M., Angenent, G. C. & Sablowski, R. Transcriptional program controlled by the floral homeotic gene AGAMOUS during early organogenesis. *Development* **132**, 429-438, doi:10.1242/dev.01600 (2005).

23 Pliner, H. A. *et al.* Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol Cell* **71**, 858-+, doi:10.1016/j.molcel.2018.06.044 (2018).

24 Gate, R. E. *et al.* Genetic determinants of co-accessible chromatin regions in activated T cells across humans. *Nat Genet* **50**, 1140-1150, doi:10.1038/s41588-018-0156-2 (2018).

25 Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat Biotechnol* **37**, 925-936, doi:10.1038/s41587-019-0206-z (2019).

26 Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486-490, doi:10.1038/nature14590 (2015).

27 Peng, Y. *et al.* Chromatin interaction maps reveal genetic regulation for quantitative traits in maize. *Nat Commun* **10**, 2632, doi:10.1038/s41467-019-10602-5 (2019).

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

28 Sun, Y. *et al.* 3D genome architecture coordinates trans and cis regulation of differentially expressed ear and tassel genes in maize. *Genome Biol* **21**, 143, doi:10.1186/s13059-020-02063-7 (2020).

29 Chuck, G. S., Brown, P. J., Meeley, R. & Hake, S. Maize SBP-box transcription factors unbranched2 and unbranched3 affect yield traits by regulating the rate of lateral primordia initiation. *Proc Natl Acad Sci U S A* **111**, 18775-18780, doi:10.1073/pnas.1407401112 (2014).

30 Heger, P., Marin, B., Bartkuhn, M., Schierenberg, E. & Wiehe, T. The chromatin insulator CTCF and the emergence of metazoan diversity. *Proc Natl Acad Sci U S A* **109**, 17507-17512, doi:10.1073/pnas.1111941109 (2012).

31 Liu, C., Cheng, Y. J., Wang, J. W. & Weigel, D. Prominent topologically associated domains differentiate global chromatin packing in rice from Arabidopsis. *Nat Plants* **3**, 742-748, doi:10.1038/s41477-017-0005-9 (2017).

32 Karaaslan, E. S. *et al.* Marchantia TCP transcription factor activity correlates with three-dimensional chromatin structure. *Nat Plants*, doi:10.1038/s41477-020-00766-0 (2020).

33 Takacs, E. M. *et al.* Ontogeny of the maize shoot apical meristem. *Plant Cell* **24**, 3219-3234, doi:10.1105/tpc.112.099614 (2012).

34 Somssich, M., Je, B. I., Simon, R. & Jackson, D. CLAVATA-WUSCHEL signaling in the shoot meristem. *Development* **143**, 3238-3248, doi:10.1242/dev.133645 (2016).

35 Jiang, K., Zhu, T., Diao, Z., Huang, H. & Feldman, L. J. The maize root stem cell niche: a partnership between two sister cell populations. *Planta* **231**, 411-424, doi:10.1007/s00425-009-1059-3 (2010).

36 Wallner, E. S. *et al.* Strigolactone- and Karrikin-Independent SMXL Proteins Are Central Regulators of Phloem Formation. *Curr Biol* **27**, 1241-1247, doi:10.1016/j.cub.2017.03.014 (2017).

37 Baker, R. F. *et al.* Sucrose Transporter ZmSut1 Expression and Localization Uncover New Insights into Sucrose Phloem Loading. *Plant Physiol* **172**, 1876-1898, doi:10.1104/pp.16.00884 (2016).

38 Kajala, K. *et al.* Innovation, conservation and repurposing of gene function in plant root cell type development. *bioRxiv*, 2020.2004.2009.017285, doi:10.1101/2020.04.09.017285 (2020).

39 Brady, S. M. *et al.* A high-resolution root spatiotemporal map reveals dominant expression patterns. *Science* **318**, 801-806, doi:10.1126/science.1146265 (2007).

40 You, Y. *et al.* Phloem Companion Cell-Specific Transcriptomic and Epigenomic Analyses Identify MRF1, a Regulator of Flowering. *Plant Cell* **31**, 325-345, doi:10.1105/tpc.17.00714 (2019).

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1284 41 Jiang, H., Li, H., Bu, Q. & Li, C. The RHA2a-interacting proteins ANAC019 and ANAC055
1285 may play a dual role in regulating ABA response and jasmonate response. *Plant Signaling*
1286 *& Behavior* **4**, 464-466, doi:10.4161/psb.4.5.8543 (2009).

1287 42 Yamaguchi, M. *et al.* VND-INTERACTING2, a NAC domain transcription factor, negatively
1288 regulates xylem vessel formation in Arabidopsis. *Plant Cell* **22**, 1249-1263,
1289 doi:10.1105/tpc.108.064048 (2010).

1290 43 Jiao, Y. *et al.* Improved maize reference genome with single-molecule technologies.
1291 *Nature* **546**, 524-527, doi:10.1038/nature22971 (2017).

1292 44 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-
1293 2079, doi:10.1093/bioinformatics/btp352 (2009).

1294 45 AFA Smit, R. H., and P Green. *RepeatMasker Open-4.0* (2013-2015).

1295 46 Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421,
1296 doi:10.1186/1471-2105-10-421 (2009).

1297 47 Team, R. C. *R: A language and environment for statistical computing.*, <http://www.R-
1298 project.org/> (2013).

1299 48 Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor.
1300 *Bioinformatics* **34**, i884-i890, doi:10.1093/bioinformatics/bty560 (2018).

1301 49 Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
1302 *arXiv e-prints*, arXiv:1303.3997 (2013).
1303 <https://ui.adsabs.harvard.edu/abs/2013arXiv1303.3997L>.

1304 50 Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing.
1305 *arXiv e-prints*, arXiv:1207.3907 (2012).
1306 <https://ui.adsabs.harvard.edu/abs/2012arXiv1207.3907G>.

1307 51 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137,
1308 doi:10.1186/gb-2008-9-9-r137 (2008).

1309 52 Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-
1310 seq data using regularized negative binomial regression. *Genome Biol* **20**, doi:ARTN 296
1311 10.1186/s13059-019-1874-1 (2019).

1312 53 Witten, D. M., Tibshirani, R. & Hastie, T. A penalized matrix decomposition, with
1313 applications to sparse principal components and canonical correlation analysis.
1314 *Biostatistics* **10**, 515-534, doi:10.1093/biostatistics/kxp008 (2009).

1315 54 Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-+,
1316 doi:10.1016/j.cell.2019.05.031 (2019).

1317 55 Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers.
1318 *Science* **362**, doi:10.1126/science.aav1898 (2018).

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.27.315499; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

56 Fang, R. *et al.* SnapATAC: A Comprehensive Analysis Package for Single Cell ATAC-seq. *bioRxiv*, 615179, doi:10.1101/615179 (2020).

57 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140, doi:10.1093/bioinformatics/btp616 (2010).

58 Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017-1018, doi:10.1093/bioinformatics/btr064 (2011).

59 O'Malley, R. C. *et al.* Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell* **166**, 1598, doi:10.1016/j.cell.2016.08.063 (2016).

60 Galli, M. *et al.* The DNA binding landscape of the maize AUXIN RESPONSE FACTOR family. *Nat Commun* **9**, 4526, doi:10.1038/s41467-018-06977-6 (2018).

61 Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat Methods* **14**, 975-978, doi:10.1038/nmeth.4401 (2017).

62 Kuhn, M. Building Predictive Models in R Using the caret Package. *2008* **28**, 26, doi:10.18637/jss.v028.i05 (2008).

63 Granja, J. M. *et al.* ArchR: An integrative and scalable software package for single-cell chromatin accessibility analysis. *bioRxiv*, 2020.2004.2028.066498, doi:10.1101/2020.04.28.066498 (2020).

64 Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 238, doi:10.1186/s13059-019-1832-y (2019).

65 van Dijk, D. *et al.* Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174**, 716-729 e727, doi:10.1016/j.cell.2018.05.061 (2018).

66 Alpert, A., Moore, L. S., Dubovik, T. & Shen-Orr, S. S. Alignment of single-cell trajectories to compare cellular expression dynamics. *Nat Methods* **15**, 267-270, doi:10.1038/nmeth.4628 (2018).

67 Hofmeister, B. T. & Schmitz, R. J. Enhanced JBrowse plugins for epigenomics data visualization. *BMC Bioinformatics* **19**, 159, doi:10.1186/s12859-018-2160-z (2018).