Evolution of cell-type-specific accessible chromatin regions and the *cis*-regulatory elements that drive lineage-specific innovation

Haidong Yan[1,+], John P. Mendieta[1,+], Xuan Zhang[1], Alexandre P. Marand[1#], Yan Liang[2], Ziliang Luo[1], Thomas Roulé[3], Doris Wagner[3], Xiaoyu Tu[4], Yonghong Wang[2,5], Silin Zhong[6], Susan R. Wessler[7], and Robert J. Schmitz[1,*]

[1]Department of Genetics, University of Georgia, Athens, GA, 30602, USA
[2]College of Life Sciences, Shandong Agricultural University, Taian, Shandong 271018, China
[3]Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA
[4]Joint Center for Single Cell Biology, School of Agriculture and Biology, Shanghai Jiao Tong University, Shanghai 200240, China
[5]Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China
[6]State Key Laboratory of Agrobiotechnology, School of Life Sciences, The Chinese University of Hong Kong, China
[7]Department of Botany and Plant Sciences, University of California, Riverside, CA, 92521, USA

[+]These authors contributed equally.
[#] Current address: Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, MI, 48109, USA
*Corresponding author

Abstract:

Cis-regulatory elements (CREs) are critical in coordinating the regulation of gene expression. Understanding how CREs evolve represents a challenging endeavor. Here, we constructed a comprehensive single-cell atlas of CRE activity in *Oryza sativa*, integrating data from 104,029 nuclei representing 128 discrete cell states across nine distinct organs. We used a comparative genomic approach to analyze cell-type-specific CRE activity between *Oryza sativa* and four additional grass species (*Zea mays, Sorghum bicolor, Panicum miliaceum*, and *Urochloa fusca*). We revealed a complex interplay whereby accessible chromatin regions (ACRs) had different levels of conservation dependent on if these regions were broadly accessible, or cell-type specific. We found that on average epidermal ACRs were less conserved in regions of synteny compared to other cell types, potentially indicating that large amounts of regulatory evolution have taken place in the epidermal layers of these species. We found a significant overlap between cell-type-specific ACRs and conserved non-coding sequences. Finally, we identified a subset of ACRs that were overlapping the repressive histone modification H3k27me3, which were conserved across evolutionary time, implicating them as potentially critical silencer CREs. Collectively, these results highlight the dynamic evolution of cell-type-specific CRE activity using comparative genomics.

Main

Cis-regulatory elements (CREs) function as pivotal hubs, facilitating the binding of transcription factors (TFs) and recruitment of chromatin-modifying enzymes, thereby fine-tuning gene expression in a spatiotemporal-specific manner[1]. CREs play important roles in developmental and environmental processes, and their functional divergence frequently serves as a key driver of evolutionary change[2]. Prior studies uncovered the prevalence of CREs in plants, highlighting their dynamic nature throughout evolution and their involvement in regulating gene expression via distinct chromatin pathways[3-8]. Across diverse cell types, gene expression is intricately regulated by multipl[9]e distinct CREs, each exerting control within specific cell or tissue types or during particular developmental stages and environmental cues[10-12]. Several studies have identified CREs functioning in a cell-type-specific manner within various plant species[9,13-19]. Despite these findings, our understanding of CREs exhibiting evolutionarily conserved or divergent cell-type-specific activities remains limited.

We initiated discovery of accessible chromatin regions (ACRs) at single-cell resolution within *Oryza sativa* making an expansive reference atlas of single-cell ACRs. We then leveraged these data in tandem with four additional scATAC libraries of leaf tissue from diverse species (*Zea mays*, *Sorghum bicolor*, *Panicum miliaceum*, and *Urochloa fusca*)[20], allowing us to compare ACRs both across species and cell types, as well as investigate whether these ACRs are dynamic across tissues. We quantified the proportion of ACRs that were conserved in this sample of monocots and found high rates of cell-type-specific ACR turnover. This indicates that the CREs associated with specific cell types are rapidly changing through evolutionary time. Finally, we used both conserved non-coding sequences (CNS), as well as epigenomic data in the form of H3K27me3 to find a series of conserved ACRs with candidate CREs within them that are important for recruitment of Polycomb-mediated gene silencing.

Construction of an ACR atlas at single-cell resolution in rice

To create a cell-type-resolved ACR atlas in rice, we conducted scATAC-seq across a spectrum of nine distinct organs in replicate (Fig. 1a and b). Evaluation of the scATAC-seq data quality, encompassing metrics such as biological replicates, transcription start site enrichment, fraction of reads in peaks, fragment size distribution, and organelle content, revealed data of excellent quality (Supplementary Fig. 1). Following strict quality control filtering, we identified 104,029 high-quality nuclei, with an average of 41,701 unique Tn5 integrations per nucleus. Upon annotation of each of the organ datasets, we identified a total of 128 cell states, encompassing 60 main cell types based on nine robust annotation strategies across various developmental stages (Fig. 1b; Extended Data Fig. 1a; Supplementary Note 1; Supplementary Fig. 2-11; Supplementary Tables 1-5).

We characterized ACRs by analyzing cell-type-aggregated chromatin accessibility profiles to identify a total of 120,048 ACRs (Extended Data Fig. 1b and c). When analyzing their proximity to genomic features in the *O. sativa* genome, we found about half of the ACRs were proximal (52%) showing higher chromatin accessibility and less variance in chromatin accessibility than genic and distal ACRs (Fig. 1c; Extended Data Fig. 1d). In contrast, we found 19% of the ACRs were genic mostly being found in intronic regions, and the remaining 29% were categorized as distal (Fig. 1c). When looking at the chromatin-accessibility variance of the non-proximal

regions, we found greater variation, potentially implicating these regions as being important in different cellular contexts, tissues or response to the environment. Among the 120,048 ACRs, 42,649 were categorized as 'cell-type-specific ACRs', exhibiting accessible chromatin in less than 17% (10/60) of the major cell types, whereas approximately 77,399 were classified as 'broad ACRs' with chromatin accessibility in more than 17% of the cell types (Fig. 1d; Extended Data Fig. 1e and f).

Although single-cell atlases are invaluable resources, we were interested in leveraging the atlas to understand how ACRs change during evolution. We overlapped this dataset to syntenic regions defined by their relationship to four different grass species *Z. mays*, *S. bicolor*, *P. miliaceum*, and *U. fusca* that have scATAC-seq data from leaves[20]. This allowed a comparison to examine ACR conservation and divergence. The analysis revealed that 34% (40,477) of the ACRs were within 8,199 syntenic regions shared between rice and at least one of four examined grass species, whereas the majority of ACRs (66%; 79,571) were located in non-syntenic regions. Notably, ~39% (30,704) of these were cell-type-specific ACRs, which were more likely to reside in non-syntenic regions (Fig. 1e; Extended Data Fig. 1g).

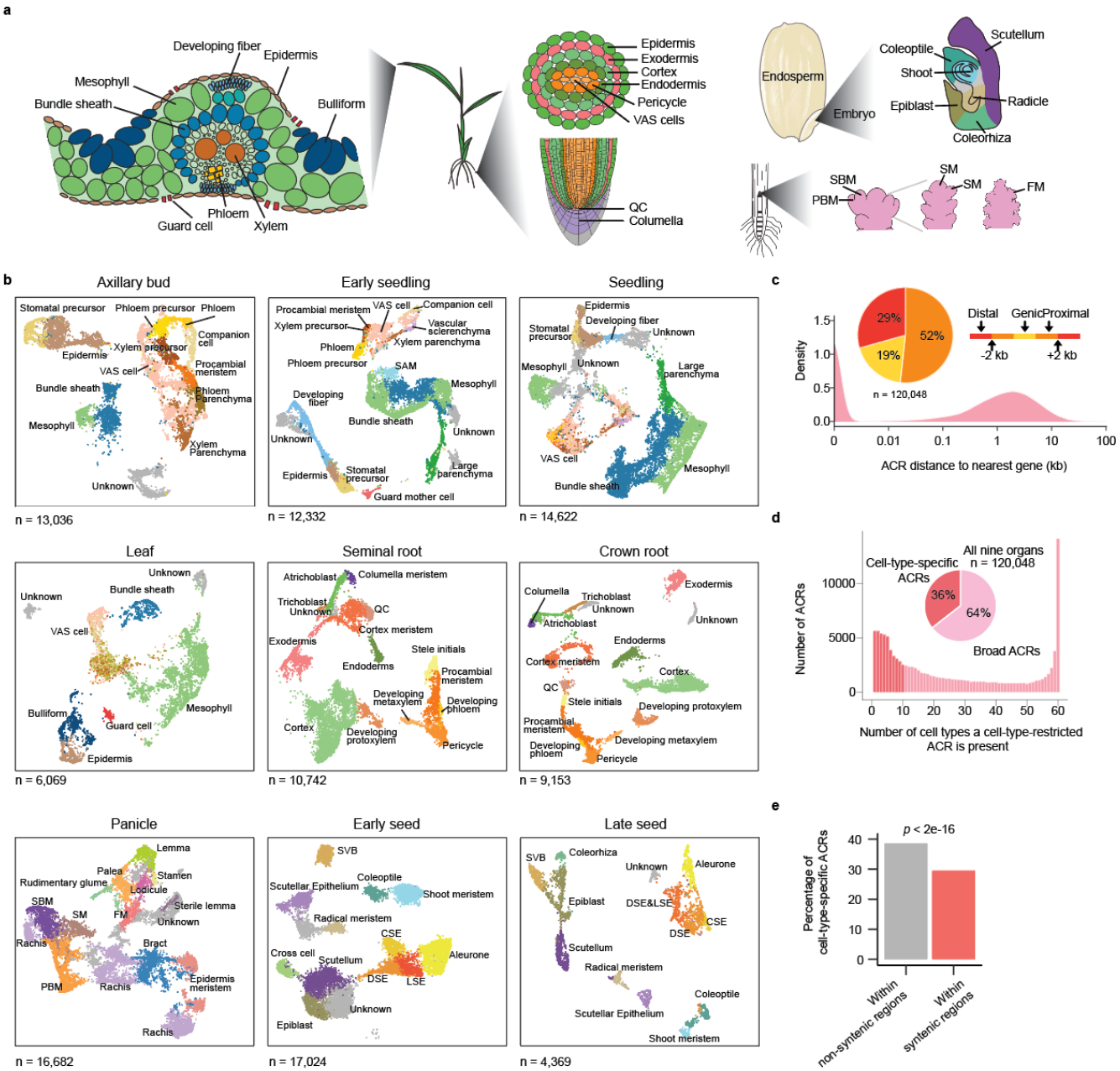


Fig. 1. Identifying cell types and characterizing ACRs in rice using scATAC-seq data. a, Overview of cell types in leaf, root, seed, and panicle organs. VAS cells: vasculature-related cells. QC: Quiescent Center. SBM: Secondary Branch Meristem. PBM: Primary Branch Meristem. SM: Spikelet Meristem. FM: Floral Meristem. **b,** UMAP projection of nuclei, distinguished by assigned cell-type labels in axillary bud, early seedling (7 days after sowing), seedling (14 days after sowing), leaf (V4 stage; four leaves with visible leaf collars), seminal root, crown root, panicle, early seed development (6 DAP; days after pollination), and late seed development (10 DAP). SAM: Shoot Apical Meristem. CSE: Central Starchy Endosperm. DSE: Dorsal Starchy Endosperm. LSE: Lateral Starchy Endosperm. SVB: Scutellar Vascular Bundle. **c,** Accessible chromatin regions (ACRs) show a bimodal distribution of distance to the nearest gene. The ACRs were categorized into three major groups based on their locations to the nearest gene: genic ACRs (overlapping a gene), proximal ACRs (located within 2 kb of genes), and distal ACRs (situated more than 2 kb away from genes). **d,** Evaluation of proportions of ACRs that are cell-type

specific versus broad. **e,** The bar plot illustrates a notable enrichment of cell-type-specific ACRs in non-syntenic regions as opposed to syntenic regions. Significance testing was performed using the Fisher's exact test (alternative = 'greater').

The landscape of cell-type-specific ACRs across species

We analyzed the composition and location of ACRs across all grass species to evaluate if there were any generalizable patterns in the number of broadly accessible (broad) and cell-type-specific ACRs found in the leaf cell types (Fig. 2a). Using the previous cell-type annotation[20], we calculated the proportion of both broad and cell-type-specific ACRs across all species. On average, ~53,000 ACRs were identified across the species, with 15-35% of the ACRs classified as cell-type specific (Fig. 2b). Broad and cell-type-specific ACRs were roughly equal in their enrichment around promoters, distal, and genic regions (Fig. 2c). Additionally, there was no pattern associated with distance to protein-coding genes when comparing cell-type-specific to broad ACRs (Extended Data Fig. 2a-c).

We used these data to associate gene-body-chromatin accessibility of specific transcription factors (TFs) with their cognate motifs, exploring this relationship across different species and cell types. Investigating motif coverage centered on ACRs (Extended Data Fig. 3a) revealed that ~71% of the TFs examined exhibited a positive correlation between the local chromatin accessibility of the TF locus and global enrichment of their cognate binding motifs within ACRs across cell types (Extended Data Fig. 3b). This finding suggests a relationship between the local TF gene chromatin accessibility and TF gene activity in the same cell. We explored for instances where divergent TF activity in orthologous cell types occurred, as it has been long implied that large scale rewiring of both TF expression and motif usage could play a key role in cell-type adaptation to variable conditions[8,21]. TF motif enrichment analysis revealed several known TF relationships within the cell types that have been previously established (Extended Data Fig. 3c). Notably, we observed predominant enrichments of WRKY, HOMEODOMAIN-LEUCINE ZIPPER (HD-ZIP), and PLANT ZINC FINGER (PLINC) in epidermis cells across all five species examined (Fig. 2d and e; Supplementary Table 6). This result indicates that these TFs might play essential roles in the development of the epidermis compared to other cell types. Interestingly, phloem companion, sieve element cells, and bundle sheath cells exhibited specific enrichment patterns in four species (*Z. mays*, *S. bicolor*, *P. miliaceum*, and *U. fusca*) that are C4 style photosynthesizers compared to *O. sativa*, which uses C3 photosynthesis (Fig. 2d; Extended Data Fig. 3d and e). One specific example that stood out was the DNA-BINDING ONE ZINC FINGER (DOF) TF family, known to act on the formation and functioning of vascular tissues in *Arabidopsis thaliana*[22], and has been implicated in the switch from C3 to C4[19,23,24] (Fig. 2d; Extended Data Fig. 3d and e). These results indicate that these TF families exhibiting distinct enrichment in phloem companion, sieve elements and bundle sheath cells might fulfill roles contributing to different vascular systems in leaves between C3 and C4 species[25,26]. These findings demonstrate the advantage of scATAC-seq data in a comparative framework to explore

genome evolution, both based on the relationship of ACRs to genomic features, as well as the relationship between TFs and their binding motifs.

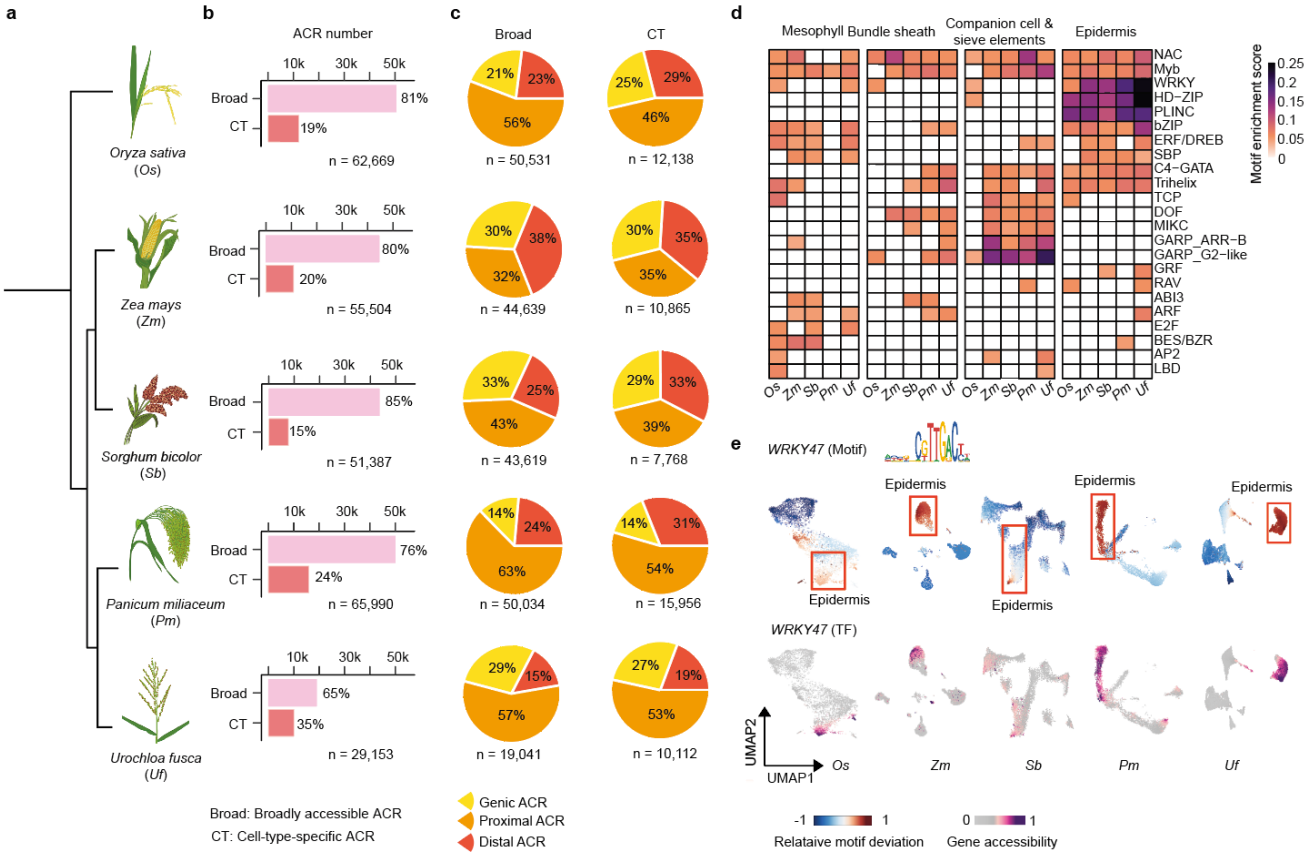


Fig. 2. Position and motif enrichment of cell-type-specific ACRs across species. a, A phylogenetic tree illustrates five species under examination. **b,** The count of broad and cell-type-specific ACRs. **c,** Broad and cell-type-specific ACRs were classified into three main groups based on their proximity to the nearest gene: genic ACRs (overlapping a gene), proximal ACRs (located within 2 kb of genes), and distal ACRs (situated more than 2 kb away from genes). *O. sativa*, *P. miliaceum*, and *U. fusca* showed a higher percentage of proximal ACRs, but a lower percentage of distal ACRs compared to *Z. mays* and *S. bicolor*, likely reflecting differences in intergenic space and overall genome sizes. **d,** A heatmap illustrates collapsed TF motif enrichment patterns into super motif families across various species for each cell type. To mitigate the impact of substantial variations in cell numbers across species or cell types, we standardized the cell count by randomly selecting 412 cells per cell type per species. This count represents the lowest observed cell count for a given cell type across all species (See Methods: Linear-model based motif enrichment analysis). The motif enrichment score cutoff was set to 0.05. **e,** Deviations of one representative TF motif underlying the WRKY family displaying enrichment consistent with the gene body chromatin accessibility of their associated TFs. The UMAP panels highlight motifs enriched with ACRs in epidermis cells, where their cognate TFs are known to accumulate.

Species-specific evolution of cell-type-specific ACRs

To understand how cell-type-specific and broad ACRs change over evolutionary time, we examined ACRs that are within syntenic regions between the studied species. To compare ACRs, we devised a pipeline that used synteny allowing us to compare sequences directly using BLASTN (See Methods: Identification of syntenic regions; Fig. 3a; Extended Data Fig. 4a; Supplementary Table 7). Using *O. sativa* ACRs as the reference, we identified 1) ACRs that match in sequences and are accessible in both species (shared ACR), 2) ACRs that match in sequence, but are only accessible in one species (variable ACR), and 3) ACRs where the sequence is exclusive to a single species (species-specific ACR; Fig. 3b). We found that, on average, the shared class of ACRs were enriched for ACRs that were broadly accessible, whereas the variable and species-specific classes were enriched in cell-type-specific ACRs (Fig. 3c; Extended Data Fig. 4b and c). Moreover, we observed that the genomic distribution of shared ACR pairs, which were predominantly broadly accessible, were located in gene-proximal regions (Extended Data Fig. 4d). This is in contrast to cell-type-specific ACRs, which are over represented in the species-specific class, and tend to reside in distal genomic regions, as compared to the broad ACRs (Extended Data Fig. 4e).

Although the conservation and position of each class of ACRs was interesting, whether the cell-type-specific ACRs were conserved in their cell-type-specificity over evolutionary time presented an interesting line of inquiry. When investigating the conservation of the cell-type-specificity of these ACRs, we found that a majority of cell-type-specific ACRs in *O. sativa* aligned with the same cell-type-specific ACRs in the corresponding species (Fig. 3d). Interestingly, we find a large proportion of companion cell and sieve element ACRs that maintain their cell-type identity. One specific example of this is the conservation of the ACR associated with the promoter of *LATERAL ROOT DEVELOPMENT 3* in *S. bicolor* and *O. sativa* (Fig. 3d), a gene critical in companion cell and sieve element development[27]. Additionally, we observed that ACRs within mesophyll and bundle sheath cells displayed shifts in their cell-type specificity (Fig. 3d), likely due to the functional divergence associated with the shift from C3 (*O. sativa*) to C4 (all other species sampled) photosynthesis[19,20].

Next, we investigated the relative conservation of cell-type-specific ACRs across the three distinct classes of ACRs. Across all the classes and species comparisons, we found that species-specific ACRs comprise the largest proportion for all the cell types (Fig. 3e). These results indicate that the cell-type-specific ACRs within these species are changing at a rapid rate. However, when evaluating the shared ACRs, we were struck by how epidermis and protoderm were the least conserved in all of the species comparisons, but they made up the highest proportion of species-specific ACRs (Fig. 3e). Therefore, we investigated the epidermal and protoderm classes of ACRs, within and outside regions of synteny to determine if there are any unique CREs associated with them. Within the species-specific ACRs within syntenic regions, we observed a predominance of CREs enriched in protoderm and epidermis cells, particularly for the HD-ZIP, SQUAMOSA PROMOTER BINDING PROTEIN (SBP), PLINC families

(Extended Data Fig. 5d and e). Many of these, such as *HDG1*, *ZHD1*, *ATHB-20*, *SPL3*, *SPL4*, and *SPL5,* are known to function in epidermal cell development[28-31]. Upon comparing TF motif enrichment in syntenic and non-syntenic ACRs, we observed the presence of these families in both groups (Extended Data Fig. 5f; Supplementary Table 8), indicating their potential indispensable roles in plant epidermal cell development. Notably, some TF-motif families such as WRKY in epidermal cells, were identified to have more TF motif enriched in non-syntenic ACRs (Extended Data Fig. 5f). These results highlight the possibility that changes in the *cis*-regulatory space associated with the epidermis might play a pivotal role associated with epidermis diversity in this sampling of monocots. Although BLASTN and comparing ACR sequences is invaluable for revealing the relative diversification and conservation of ACRs, computational methods aimed at identifying CNS in plants have been used as an important proxy to identify functional elements *in lieu* of accessible chromatin data[32-34].

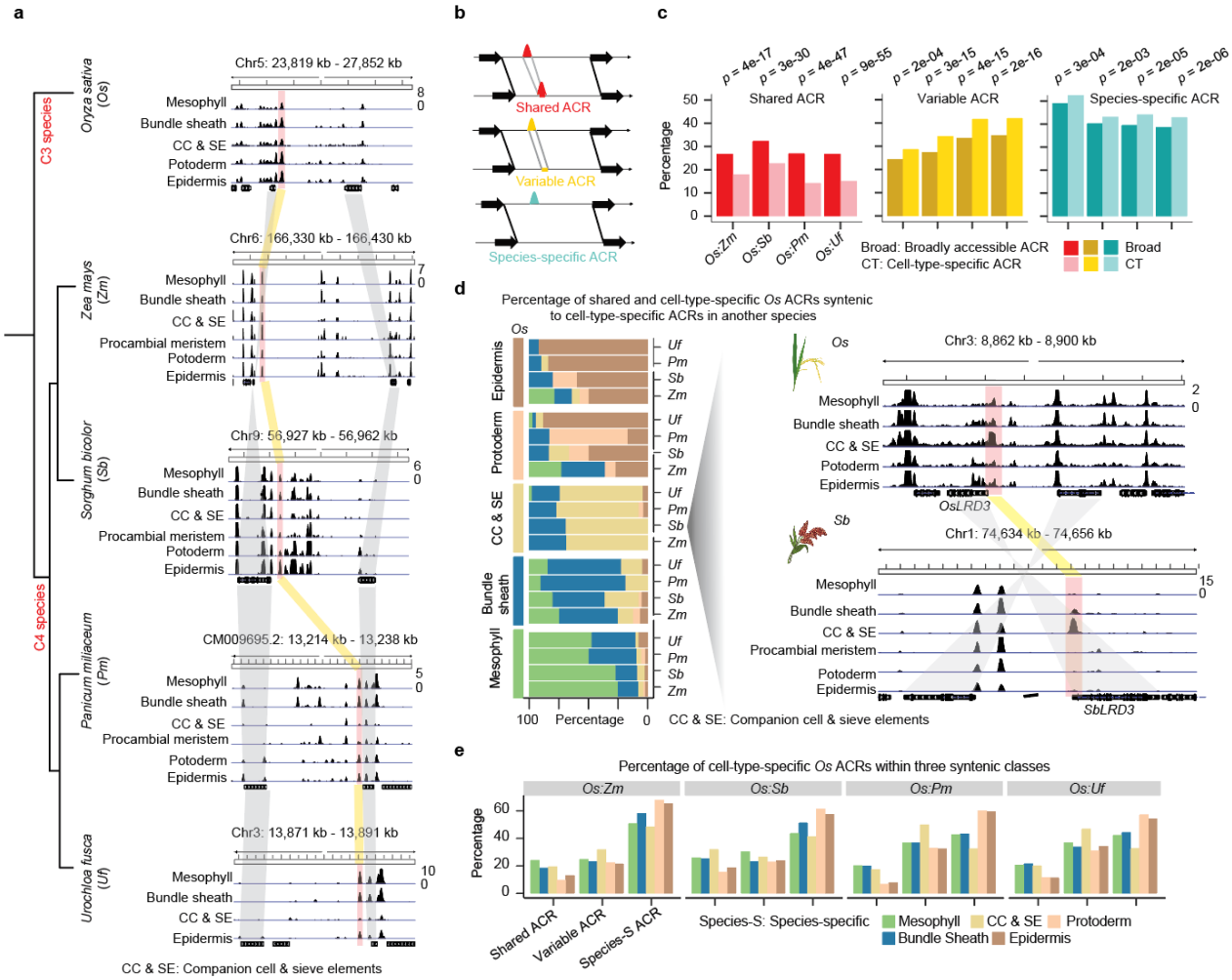


Fig. 3. Cell-type-specific ACRs are frequently species-specific. a, A screenshot of syntenic regions capturing shared ACRs across five species. The red bars denote syntenic ACRs within regions flanked by corresponding syntenic gene pairs, while the gray color highlights these syntenic gene pairs. **b,** Three

classes depicting variations in ACR conservation between two species. **c,** The percentage of broad and cell-type-specific ACRs underlying three classes shown in panel b. The significance test was done by using the Fisher's exact test (alternative = 'greater'). **d,** The percentage of shared and cell-type-specific ACRs among various cell types in four species, which are syntenic to the cell-type-specific and shared ACRs in *O. sativa*. The right panel showed a screenshot of an ACR in *O. sativa* that is specific to companion cell and sieve elements, which is syntenic to a corresponding ACR specific to the same cell type in *S. bicolor*. This ACR is situated at the promoter region of a marker gene *LRD3* specifically expressed in companion cell and phloem sieve elements (Supplementary Table 2). The gray shaded regions indicate a syntenic gene pair. **e,** The percentage of cell-type-specific ACRs identified collectively across all cell types within each species pair under three syntenic classes. The percentage for each cell type within the three classes collectively sum to 100%.

CNS are enriched in cell-type-specific ACRs

A common method to identify candidate CREs in the absence of accessible chromatin data is to use a comparative genomics approach in search of CNS[6,35-37]. Using published CNS[33], excluding those overlapping with untranslated regions, we observed that 30.8% and 21.3% of CNS overlapped with ACRs identified in the leaf cell types in *O. sativa* and *Z. mays*, respectively (Fig. 4a). Using ACRs from the *O. sativa* atlas increased this ratio to 65.0% (Fig. 4a), indicating that a significant portion of CNS are captured by ACRs. Notably, ACRs that overlapped with CNS within syntenic regions are more likely cell-type-specific ACRs (Fig. 4b and c), suggesting the CNS have been maintained over evolutionary time are potentially important in governing specific cellular activities. Interestingly, we observed that CNS from *Z. mays* that overlapped leaf cell-type-specific ACRs were commonly found in non-leaf cell types when assessed against the rice atlas (Fig. 4d and e). This suggests co-option of CREs into variable tissues or cell types, likely caused by either nearby CRE evolution or *trans* factors that have evolved to provide this new cell-type specificity. Lastly, epidermis cells showed a significantly lower ratio of species-specific ACRs capturing CNS compared to other cell types (Fig. 4f), consistent with our previous findings that epidermal cell-type-specific ACRs tend to be associated with species-specific ACRs (Fig. 3e; Extended Data Fig. 5a-c).

Interestingly, we noticed a pattern where many CNS overlapping syntenic ACRs also overlapped domains of the repressive histone modification H3K27me3. In brief, H3K27me3 is a facultative mark of heterochromatin established by the POLYCOMB REPRESSIVE COMPLEX 2 (PRC2)[38-40]. Genes targeted by PRC2 and H3K27me3 are important developmental loci that are only expressed during distinct stages of development, making the identification of key CREs important for silencing especially interesting[39,41]. When comparing ACRs overlapping domains of H3K27me3, we found that these ACRs were statistically enriched in CNS, potentially indicating that some of these CNS are acting as silencers (Fig. 4g and h).

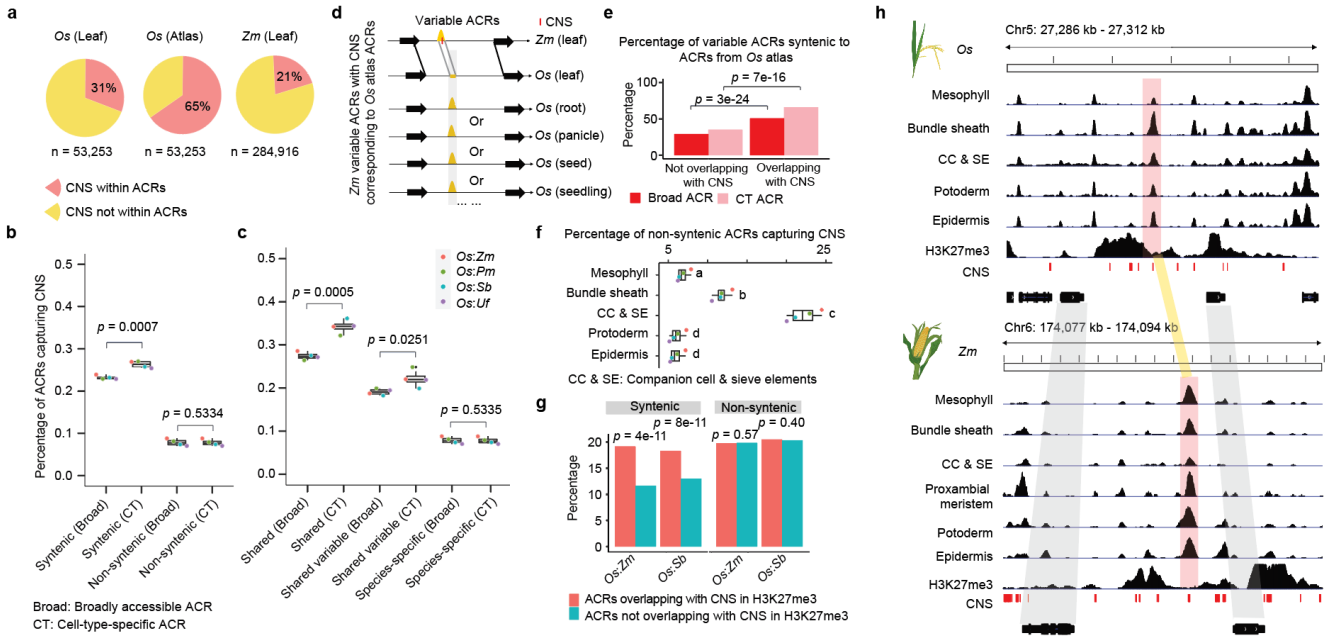


Fig. 4. Cell-type-specific ACRs exhibit an enrichment of CNS. a, The percentage of CNS overlapping with ACRs. 'n' indicates the number of CNS. 'Atlas' means the ACRs were from the built *O. sativa* atlas in Fig. 1c. **b,** The percentage of broad and cell-type-specific ACRs within syntenic and non-syntenic regions overlapping with the CNS. Significance testing was performed using the t-test (alternative = 'greater'). **c,** The box plot presented similar meaning as panel b but focusing on three classes within syntenic regions shown in Fig. 3b. **d,** A sketch graph illustrates whether variable ACRs containing CNS in *Z. mays* capture ACRs derived from the *O. sativa* atlas. **e,** The bar plot depicts the percentage of variable ACRs captured in the *O. sativa* atlas. Significance testing was performed using the Fisher's exact test (alternative = 'greater'). **f,** The percentage of non-syntenic ACRs capturing CNS. All pairwise comparisons are statistically significant ($p < 0.05$) as indicated by different letters based on t-test (alternative = 'greater'). If two bars have the same letter, then they are not significantly ($p > 0.05$) different from each other. **g,** The percentage of ACRs capturing CNS within and not within H3K27me3 regions. The percentage for each group within H3K27me3 and not within H3K27me3 regions collectively sum to 100%. Significance testing was performed using the Fisher's exact test (alternative = 'greater'). **h,** An example of a syntenic region containing ACRs within the H3K27me3 regions that were conserved between *O. sativa* and *Z. mays*. CNS were highlighted using red color.

Candidate silencer CREs are enriched in broad ACRs

To assess candidate silencer ACRs across diverse species, we focused on scATAC-seq and H3K27me3 data from *O. sativa*, *Z. mays*, and *S. bicolor*. We identified ACRs that overlap or directly flank H3K27me3 regions and classified them based on their proximity to H3K27me3 peaks and their frequency across cellular contexts (Fig. 5a). Notably, broad ACRs exhibited a depletion of H3K27me3, consistent with the reduced presence of nucleosomes in ACRs. In contrast, depletion of H3K27me3 was not observed in the cell-type-specific ACRs that overlapped H3K27me3 regions (Fig. 5b). The composition of H3K27me3-broad and cell-type-specific ACRs was consistent across all species, with the majority of H3K27me3-broad ACRs being either proximal or distal to their target gene (Extended Data Fig. 6a).

We hypothesized that many of the H3K27me3-broad ACRs are enriched for silencer elements, as these regions are accessible in most cell types to maintain H3K27me3 throughout development to silence gene expression. Using snRNA-seq data from *O. sativa* and *Z. mays* seedling and root tissues, we found significantly lower expression ($p < 0.05$) for genes associated with H3K27me3-ACRs per cell type (Fig. 5c; Extended Data Fig. 6b and c; Supplementary Fig. 6; Supplementary Table 10). Moreover, using 58 bulk RNA-seq libraries from seedlings, roots, and panicles in *O. sativa*[42], we found that the average expression levels of genes located near H3K27me3-broad ACRs in one organ were significantly lower than the expression levels of genes that possess H3K27me3-absent ACRs in the other organs (Extended Data Fig. 6d). To dissect roles of the H3K27me3-broad ACRs, we identified a total of 2,166 H3K27me3-broad ACRs and examined the expression profiles of their neighboring genes from *O. sativa* seedlings (Supplementary Table 11). Approximately 42.6% (923/2,166) of these ACRs surrounding 836 genes exhibited no expression across any cell type (Extended Data Fig. 6e), suggesting a potential role as constitutive silencers. For the remaining 1,243 ACRs associated with expressed genes, about 65% exhibited a weak correlation ($|rho| < 0.4$) between chromatin accessibility and gene expression. Interestingly, about 116 ACRs in this group demonstrated cell-type-specific gene expression, suggesting a potential dual enhancer function in a specific cell type. Moreover, ~11.3% (141/1,243) of the ACRs displayed high anti-correlation ($rho <= -0.4$), whereas 23.3% (289/1,243) exhibited positive correlation ($rho >= 0.4$) with nearby gene expression. This indicates that the positively correlated ACRs might coordinate with other CREs in a collective manner to regulate gene expression. In summary, these findings suggest the presence of H3K27me3-broad ACRs likely harbor silencer CREs important for affecting expression of nearby genes.

To provide additional support for candidate silencer CREs, we re-analyzed scATAC-seq data from *A. thaliana* roots[16]. Our findings revealed that H3K27me3-broad ACRs significantly ($p <$ 2e-16) captured around one-third of the 170 known Polycomb responsive elements identified from a previous study, supporting that they likely function as silencers[39] (Extended Data Fig. 6f). Furthermore, we performed *de novo* motif analysis on the 170 elements found in *A. thaliana*, which identified all six known Polycomb responsive element motifs (CTCC, CCG, G-box, GA-repeat, AC-rich, and Telobox) (Fig. 5d). Using these motifs, we predicted their binding sites in *O. sativa*, *Z. mays*, and *S. bicolor*, and observed that five motifs were significantly enriched in the H3K27me3-broad ACRs (Fig. 5d). We further observed that the majority of H3K27me3-broad ACRs contained at least one of the six motifs known to function as Polycomb responsive elements, with the majority harboring more than one type, and a few encompassing all the six (Extended Data Fig. 6g). Additionally, EMF2b is one of core components in PRC2 in *O. sativa* and is essential for PRC2-mediated gene silencing[39,43]. We observed 63.7% (4,053/6,364) of the H3K27me3-broad ACRs were significantly ($p <$ 2e-16) captured by EMF2b genomic targets in *O. sativa*[44] (Fig. 5e). EMF2b was enriched at the center of H3K27me3-broad ACRs (Fig. 5f), supporting the hypothesis that these ACRs recruit PRC2 to their targets. Next, we observed significant enrichment of four TF families in the H3K27me3-broad ACRs, namely, the APETALA2-like (AP2), basic Helix-Loop-Helix (bHLH), Basic Leucine Zipper (bZIP), and C2H2 zinc-finger (ZnF) family (Extended Data Fig. 6h; Supplementary Table 12). The AP2 and C2H2 were previously reported to recruit PRC2[39]. Finally, we identified a H3K27me3-broad ACR containing a known silencer region located around 5.3 kb upstream of the *FRIZZY*

PANICLE gene in rice[45] (Extended Data Fig. 6i). In general, these results lead to support that H3K27me3-broad ACRs potentially harbor H3K27me3-mediated silencers.

To further investigate the relationship between candidate silencers and species divergence, we classified the *O. sativa* ACRs into three classes according to the syntenic ACRs identified between the two species within H3K27me3 regions (Extended Data Fig. 7a and b; Supplementary Table 13). The results revealed that H3K27me3-broad ACRs were more likely to be species-specific ACRs compared to no-H3K27me3-broad class (Extended Data Fig. 7c), supporting the hypothesis that differences in putative silencing activity among species may also contribute to species divergence (Extended Data Fig. 5h). Moreover, we observed that variable H3K27me3-broad ACRs examined in leaf cell types were commonly found in non-leaf cell types when assessed against the rice atlas (Extended Data Fig. 7d and e). This suggests a tendency for the candidate silencer CREs having evolved to contribute to this new cell-type specificity. Collectively, these observations highlight our data as a valuable resource for identification of candidate CRE activity across different species and tissues, particularly emphasizing a potential association between candidate silencer CREs and evolutionary divergence across species.

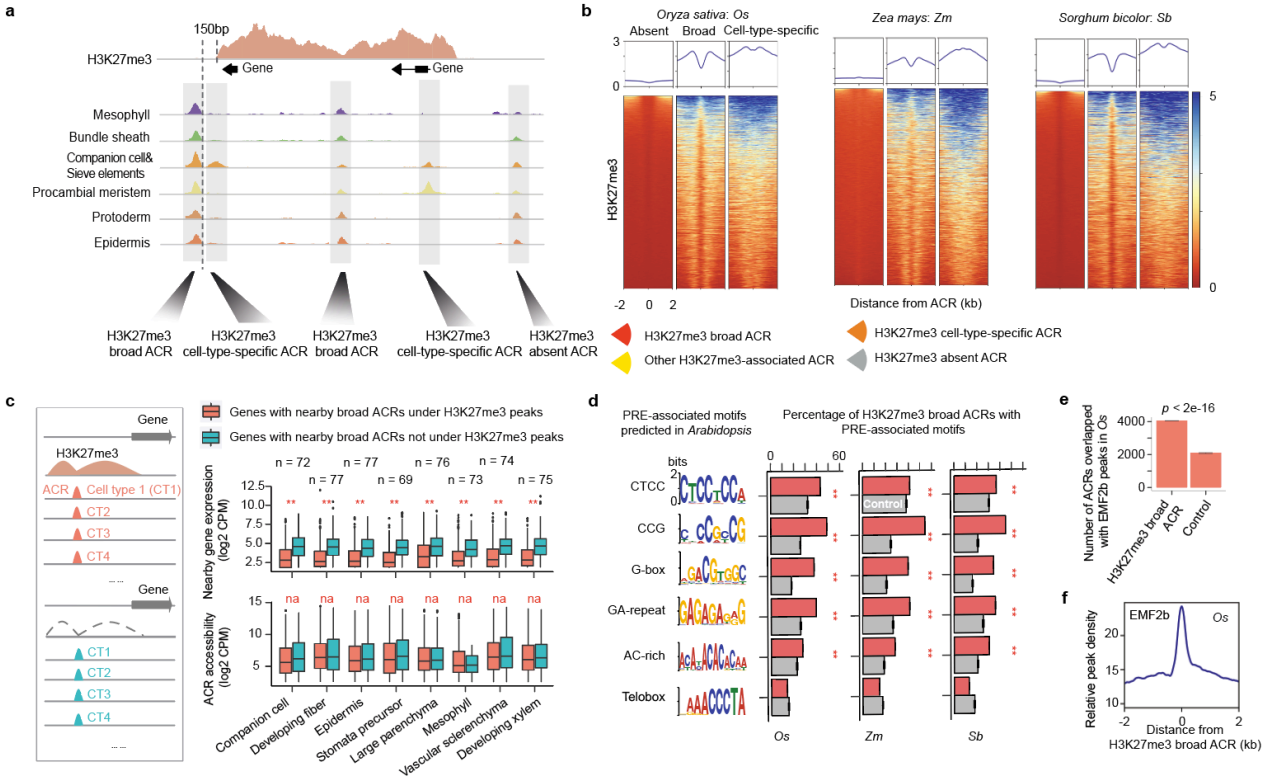


Fig. 5. Discovery of candidate silencer CREs across species. a, A sketch graph illustrates the classification of ACRs based on their proximity to H3K27me3 peaks. We classified the ACRs into two groups: H3K27me3-associated ACRs (found within or surrounding H3K27me3 peaks) and H3K27me3-absent ACRs. The H3K27me3-associated ACR were further divided into broad ACRs, characterized by chromatin accessibility in at least five cell types, and cell-type-specific ACRs, accessible in less than three out of six examined cell types across all the species. **b,** Alignment of H3K27me3 chromatin attributes at summits of distinct ACR groups. **c,** A comparative analysis of expression levels and

chromatin accessibility of genes surrounding broad ACRs under and outside of H3K27me3 peaks. ** indicate *p* value < 0.01, which was performed by the Wilcoxon test. The broad ACRs where the H3K27me3 region overlapped >50% of the gene body were positioned within 500 to 5,000 bp upstream of the transcriptional start site of their nearest gene. **d,** Percentage of H3K27me3-broad ACRs in *O. sativa*, *Z. mays*, and *S. bicolor* capturing six known motifs enriched in PREs in *A. thaliana*. **e**, The number of H3K27me3-broad-ACR capturing EMF2b ChIP-seq peak. **f**, A metaplot of EMF2b ChIP-seq profile at summits of H3K27me3-broad ACRs identified from *O. sativa* leaf.

Discussion

Understanding how ACRs and the CREs within them change over evolutionary time is an important and challenging subject. We used a comparative epigenomics approach to identify and characterize ACRs at single-cell resolution. We demonstrate that cell-type-specific ACRs across numerous grass species are changing rapidly, with relatively few cell-type-specific ACRs remaining conserved across all five species. Interestingly, of those which are conserved, we find they are overwhelmingly found in vasculature related cells (companion cells and sieve elements). This is consistent with a previous result in *Z. mays* where vasculature related cell types had less signatures of positive selection on their corresponding ACRs[16], potentially indicating that many vasculature-related cell types are of critical importance and are highly maintained. Additionally, we found a lack of conservation of epidermal-specific ACRs, likely reflecting the rapid rate of regulatory evolution of the epidermis cells. This is possible due to the constant evolutionary pressures the epidermis is exposed to, leading to rapid alteration of *cis*-regulatory sequences[46].

The use of the *O. sativa* cell-type ACR atlas allowed us to identify a set of ACRs present within H3K27me3 regions across the majority of cell types. We hypothesize these are putative silencer *cis*-regulatory regions, as they were also enriched for CNS, PRC2 binding and low gene expression of nearby genes. The epigenomic landscape of broad ACRs associated with the putative silencer regions was also present across all five grass species suggesting this is a conserved feature of flowering plant genomes. Future genome-editing studies to delete or augment these putative silencer sequences will be required to better understand their role in PRC2 recruitment and gene silencing.

Integration of the *O. sativa* atlas dataset with CNS revealed a significant number (~65%) of previously identified CNS. This highlights the need to continue assembly of reference plant genomes and use them in CNS analysis. However, it also highlights the need for additional accessible chromatin datasets from unique tissues and environments. Only with comprehensive identification of ACRs will we know the bounds of CNS analysis. Additionally, our results do point to tissue and cell-type-specific switching of CRE activity associated with CNS between species. This demonstrates that although CNS are a valuable proxy for conserved non-coding function, they don't necessarily maintain the same function across evolution[47]. The rice atlas of cell-type-specific ACRs and the cross-species analysis provides a useful resource to significantly enhance our understanding of evolution of CREs across plant species, identify the key CREs important for gene regulation and demarcate high-confident targets for improving crop performance using genome editing. .

Methods

Preparation of plant materials

Early seedlings, specifically seedling tissues above ground, were collected 7 and 14 days after sowing. Flag leaf tissue was harvested at the V4 stage, characterized by collar formation on leaf 4 of the main stem. Axillary buds were obtained from rice plants grown in the greenhouse at approximately the V8 stage. Rice seminal and crown root tips (bottom 2 cm) were gathered at the same stage as seedling tissues, 14 days after sowing. Panicle tissue was acquired from rice plants grown in the greenhouse. Inflorescence primordia (2-5 mm) were extracted from shoots harvested at the R1 growth stage, where panicle branches had formed. Early seeds were harvested at approximately six days after pollination (DAP), and late seeds at approximately ten DAP. All tissues were collected between 8 and 9 am, and all fresh materials were promptly utilized for library construction starting at 10 am.

Single-cell ATAC-seq library preparation

Nuclei isolation and purification were performed as described previously[48]. In brief, the tissue was finely chopped on ice for approximately 2 minutes using 600 µL of pre-chilled Nuclei Isolation Buffer (NIB, 10 mM MES-KOH at pH 5.4, 10 mM NaCl, 250 mM sucrose, 0.1 mM spermine, 0.5 mM spermidine, 1 mM DTT, 1% BSA, and 0.5% TritonX-100). After chopping, the entire mixture was passed through a 40-µm cell strainer and then subjected to centrifugation at 500 rcf for 5 minutes at 4°C. The supernatant was carefully decanted, and the pellet was reconstituted in 500 µL of NIB wash buffer, which consisted of 10 mM MES-KOH at pH 5.4, 10 mM NaCl, 250 mM sucrose, 0.1 mM spermine, 0.5 mM spermidine, 1 mM DTT, and 1% BSA. The sample was filtered again, this time through a 10-µm cell strainer, and then gently layered onto the surface of 1 mL of a 35% Percoll buffer, prepared by mixing 35% Percoll with 65% NIB wash buffer, in a 1.5-mL centrifuge tube. The nuclei were subjected to centrifugation at 500 rcf for 10 minutes at 4°C. Following centrifugation, the supernatant was carefully removed, and the pellets were resuspended in 10 µL of diluted nuclei buffer (DNB, 10X Genomics Cat# 2000207). About 5 µL of nuclei were diluted 10 times and stained with DAPI (Sigma Cat. D9542) and then the nuclei quality and density were evaluated with a hemocytometer under an epifluorescence microscope. The original nuclei were diluted with a DNB buffer to a final concentration of 3,200 nuclei per uL. Finally, 5 uL of nuclei (16,000 nuclei in total) were used as input for scATAC-seq library preparation. scATAC-seq libraries were prepared using the Chromium scATAC v1.1 (Next GEM) kit from 10X Genomics (Cat# 1000175), following the manufacturer's instructions. (10xGenomics, CG000209_Chromium_NextGEM_SingleCell_ATAC_ReagentKits_v1.1_UserGuide_RevE). Libraries were sequenced with Illumina NovaSeq 6000 in dual-index mode with eight and 16 cycles for i7 and i5 index, respectively.

Single-nuclei RNA-seq library preparation and data analysis

The protocol for nuclei isolation and purification was adapted from the previously described scATAC-seq method. In summary, to minimize RNA degradation and leakage, the tissue was finely chopped on ice for approximately 1 minute using 600 µL of pre-chilled Nuclei Isolation Buffer containing 0.4U/µL RNase inhibitor (Roche, Protector RNase Inhibitor, Cat. RNAINH-RO) and a comparatively low detergent concentration of 0.1% NP-40. Following chopping, the entire mixture was passed through a 40-µm cell strainer and then subjected to centrifugation at

500 rcf for 5 minutes at 4°C. The supernatant was carefully decanted, and the pellet was reconstituted in 500 µL of NIB wash buffer, comprising 10 mM MES-KOH at pH 5.4, 10 mM NaCl, 250 mM sucrose, 0.5% BSA, and 0.2U/µL RNase inhibitor. The sample was filtered again, this time through a 10-µm cell strainer, and gently layered onto the surface of 1 mL of a 35% Percoll buffer. The Percoll buffer was prepared by mixing 35% Percoll with 65% NIB wash buffer in a 1.5-mL centrifuge tube. The nuclei were then subjected to centrifugation at 500 rcf for 10 minutes at 4°C. After centrifugation, the supernatant was carefully removed, and the pellets were resuspended in 50 µL of NIB wash buffer. Approximately 5 µL of nuclei were diluted tenfold and stained with DAPI (Sigma Cat. D9542). Subsequently, the nuclei's quality and density were evaluated with a hemocytometer under a microscope. The original nuclei were further diluted with a DNB buffer to achieve a final concentration of 1,000 nuclei per µL. Ultimately, a total of 16,000 nuclei were used as input for snRNA-seq library preparation. For snRNA-seq library preparation, we employed the Chromium Next GEM Single Cell 3'GEM Kit v3.1 from 10X Genomics (Cat# PN-1000123), following the manufacturer's instructions (10xGenomics, CG000315_ChromiumNextGEMSingleCell3-_GeneExpression_v3.1_DualIndex_RevB). The libraries were subsequently sequenced using the Illumina NovaSeq 6000 in dual-index mode with 10 cycles for the i7 and i5 indices, respectively.

The raw BCL files obtained after sequencing were demultiplexed and converted into FASTQ format using the default settings of the 10X Genomics tool cellranger make-fastq[49] (v7.0.0). The raw reads were processed with cellranger count[49] (v7.0.0) using the Japonica rice reference genome[50] (v7.0). Genes were kept if they were expressed in more than three cells and each cell having a gene expression level of at least 1,000 but no more than 10,000 expressed genes. Cells with over 5% mitochondria or chloroplast counts were filtered out. The expression matrix was normalized to mitigate batch effects based on global-scaling normalization and multicanonical correlation analysis in Seurat[51] (v4.0). The Scrublet tool[52] was employed to predict doublet cells in this dataset. SCTransform in Seurat was used to normalize the data and identify variable genes. The nearest neighbors were computed using FindNeighbors using 30 PCA dimensions. The clusters were identified using FindClusters with a resolution of 1. The cell types were annotated based on the marker gene list (Supplementary Table 2).

Slide-seq library preparation and data analysis

Root tissues from rice seedlings 14 days after sowing were used for the Slide-seq V2 spatial transcriptomics. The tissues were embedded in the Optimal Cutting Temperature (OCT) compound, snap-frozen in a cold 2-methylbutane bath, and cryosectioned into 10 um thick slices. The spatial transcriptome library was constructed following the method[53,54]. In brief, the tissue slices were placed on the Slide-seq V2 puck and underwent the RNA hybridization and reverse transcription process. After tissue clearing and spatial bead collections, the cDNA was synthesized and amplified for a total of 14 cycles. The library was constructed using Nextera XT Library Prep Kit (Illumina, USA) following the manufacturer’s instructions.

The reads alignment and quantification were conducted following the Slide-seq pipeline (https://github.com/MacoskoLab/slideseq-tools). The data processing is similar to the procedures applied in the analysis of snRNA-seq but setting a resolution of 0.7 for the FindClusters function in Seurat[51] (v4.0). The cell types were annotated based on the histology of cross-sectioned roots.

The marker genes of each cluster were identified using the Wilcoxon Rank Sum test in FindAllMarkers.

RNA *in situ* Hybridization

The rice samples were put into the vacuum tissue processor (HistoCore PEARL, Leica) to fix, dehydrate, clear, and embed, and were subsequently embedded in paraffin (Paraplast Plus, Leica). The samples were sliced into 8-μm sections with a microtome (Leica RM2265). The cDNAs of the genes were amplified with their specific primer pairs in situ_F/in situ_R and subcloned into the pGEM-T vector (Supplementary Table 14). The pGEM-gene vectors were used as the template to generate sense and antisense RNA probes. Digoxigenin-labelled RNA probes were prepared using a DIG Northern Starter Kit (Roche) according to the manufacturer's instructions. Slides were observed under bright fields through a microscope (ZEISS) and photographed with an Axiocam 512 color charge-coupled device (CCD) camera.

Raw reads processing of scATAC-seq

Data processing was executed independently for each tissue and/or replicate. Initially, raw BCL files were demultiplexed and converted into FASTQ format, utilizing the default settings of the 10X Genomics tool cellranger-atac make-fastq[55] (v1.2.0). Employing the Japonica rice reference genome[50] (v7.0), the raw reads underwent processing using cellranger-atac count[55] (v1.2.0). These steps encompassed adaptor/quality trimming, mapping, and barcode attachment/correction. Subsequently to the initial processing, reads that were uniquely mapped with mapping quality > 10 and correctly paired were subjected to further refinement through SAMtools view[56] (v1.7; -f 3 -q 10). This refined set of reads was additionally filtered based on 'XA' tags. To mitigate the impact of polymerase chain reaction (PCR) duplicates, the Picard tool MarkDuplicates[57] (v2.16.0) was applied on a per-nucleus basis. To elevate data quality, a blacklist of regions was devised to exclude potentially spurious reads. The methodology involved the exclusion of regions displaying bias in Tn5 integration from Tn5-treated genomic DNA. Specifically, regions characterized by 1-kilobase windows with coverage exceeding four times the genome-wide median were eliminated. We further leveraged ChIP-seq input data from Lu et al. 2019[6], collapsed sequences in the reference were filtered out using the same criteria. This blacklist also incorporated sequences of low-complexity and homopolymeric sequences through RepeatMasker[58] (v4.1.2). Moreover, nuclear sequences exhibiting homology surpassing 80% to mitochondrial and chloroplast genomes[59] (BLAST+; v2.11.0) were also included within the blacklist. Furthermore, BAM alignments were converted into BED format, wherein the coordinates of reads mapping to positive and negative strands were subjected to a shift by +4 and -5, respectively. The unique Tn5 integration sites per barcode were finally retained for subsequent analyses.

Identifying high-quality nuclei

To ensure the acquisition of high-quality nuclei, we harnessed the capabilities of the Socrates package[16] for streamlined processing. To gauge the fraction of reads within peaks, we employed MACS2[60] (v2.2.7.1) with specific parameters (genomesize = 3e8, shift= -75, extsize = 100, fdr = 0.05) on the bulk Tn5 integration sites. Subsequently, we quantified the number of integration sites per barcode using the callACRs function. Next, we estimated the proximity of Tn5 integration sites to genes, focusing on a 2 kb window surrounding the TSS. This estimation was achieved through the buildMetaData function, which culminated in the creation of a meta file.

For further refinement of cell selection, we harnessed the findCells function, implementing several criteria: 1) A minimum read depth of 1,000 Tn5 integration sites was required. 2) The total number of cells was capped at 16,000. 3) The proportion of reads mapping to TSS sites was above 0.2, accompanied by a z-score threshold of 3. 4) Barcode FRiP scores were required to surpass 0.1, alongside a z-score threshold of 2. 5) We filtered out barcodes exhibiting a proportion of reads mapping to mitochondrial and chloroplast genomes that exceeded two standard deviations from the library mean. 6) We finally used the detectDoublets function to estimate doublet likelihood by creating synthetic doublets and conducting enrichment analysis. These multiple steps ensured the meticulous identification and selection of individual cells, facilitating a robust foundation for subsequent analyses.

Nuclei clustering

For the nuclei clustering, we leveraged all functions from the Socrates package[16]. We binned the entire genome into consecutive windows, each spanning 500 bp. We then tabulated the count of windows featuring Tn5 insertions per cell. Barcodes falling below one standard deviation from the mean feature counts (with a Z-score less than 1) were excluded. Moreover, barcodes with fewer than 1,000 features were eliminated. We pruned windows that exhibited accessibility in less than 0.5% or more than 99.5% of all nuclei. To standardize the cleaned matrix, we applied a Term Frequency-Inverse Document Frequency normalization function. The dimensionality of the normalized matrix underwent reduction through the utilization of Non-Negative Matrix Factorization (NMF), facilitated by the R package RcppML[61] (v0.3.7). We retained 50 column vectors from an uncentered matrix. Subsequently, we selected the top 30,000 windows that displayed the highest residual variance across all cells. This selection was based on fitting a model where the proportion of cells with accessibility served as an independent variable and variance as the dependent variable. To further reduce the dimensionality of the nuclei embedding, we employed the UMAP technique using umap-learn (k = 50, min_dist = 0.1, metrix = 'euclidean') in R[62] (v0.2.8.0). Furthermore, we clustered nuclei using the callClusters function within the Socrates framework[16]. Louvain clustering was applied, with a setting of k=50 nearest neighbors at a resolution of 0.7. This process underwent 100 iterations with 100 random starts. Clusters with an aggregated read depth of less than 1 million and 50 cells were subsequently eliminated. To filter outlying cells in the UMAP embedding, we estimated the mean distance for each nucleus using its 50 nearest neighbors. Nuclei that exceeded 3 standard deviations from the mean distance were deemed outliers and removed from consideration.

Estimation of gene accessibility scores

To estimate the gene accessibility, we employed a strategy wherein the Tn5 insertion was counted across both the gene body region as well as a 500 bp extension upstream. Subsequently, we employed the SCTransform algorithm from the Seurat package[51] (v4.0) to normalize the count matrix that was then transformed into normalized accessibility score, with all positive values scaled to 1.

Cell type validation

Upon completing the initial annotation process, which was based on a curated list of marker genes, we further expanded our marker repertoire by incorporating markers collected from published bulk RNA-seq data encompassing a diverse array of cell types, which were acquired *via* laser capture dissection. In brief, we collected the markers from several studies, including

three distinct cell types within rice leaves[63] and ten cell types across rice seed organs[64,65]. From these sources, we selected the top 100 variably expressed markers for each cell type and employed them to compute cell identity enrichment scores. We undertook a comprehensive assessment of markers linked to the different cell types. Subsequently, we randomly drew 100 markers from this pool and repeated this procedure 1,000 times to construct a null distribution based on marker chromatin accessibility scores. For each target cell type marker, we compared their accessibility scores to this null distribution. This facilitated the derivation of an enrichment score per cell, delineating the marker's significance for each representative cell type. We next employed a MAGIC algorithm[66] to refine these enrichment scores. These scores were then mapped onto a UMAP plot, enhancing the cell identity annotation.

Furthermore, we undertook validation of our single-cell chromatin accessibility atlas through integration with published scRNA-seq from rice root tissue. This validation was achieved through two distinct approaches (Supplementary Fig. 6). In the first approach, we leveraged the marked enrichment technique, adapting the above mentioned methodology with the incorporation of the top 20 markers derived from marker identification using the 'FindMarkers' function in Seurat[51] (v4.0). Following the acquisition of a smoothed score for each cell type, individual cells were annotated to specific cell types based on the largest enrichment score within that cell type. A threshold was further set, requiring the maximum score to exceed 0.5 for confident labeling; otherwise, the cell was labeled as 'Unknown'. The second approach entailed employing a k-nearest neighbor (knn) strategy. This strategy commenced with the normalization of scRNA-seq datasets, mirroring the process applied to scATAC-seq datasets. The top 3,000 most variable genes within the scATAC-seq dataset were then identified using the Seurat function 'FindVariableFeatures', subsequently filtering to include only genes common to both datasets. By treating the scRNA-seq cells as a reference, a dimension reduction process was conducted to generate a loading matrix, which was then utilized to project the scATAC-seq cells onto the scRNA-seq cell embedding. The integration of these two datasets was achieved through the Harmony algorithm[67] (v0.1.0). Within the dual embeddings, the 20 nearest neighbors of each scATAC-seq cell in the scRNA-seq dataset were computed. The most frequent label among these RNA neighbors (> 10 cells) was subsequently assigned as the label for each scATAC-seq cell or designated as NA if no label meeting this threshold was identified.

Cell cycle prediction

The prediction of cell cycle stages per nucleus was executed similarly to annotating cell identities based on the aforementioned enriched scores. In brief, we collected a set of 55 cell-cycle marker genes from Pettkó-Szandtner et al. 2015[68]. For every cell-cycle stage, the cumulative gene accessibility score for each nucleus was computed. These resultant scores were subsequently normalized using the mean and standard deviation derived from 1,000 permutations of the 55 random cell-cycle stage genes, with exclusion of the focal stage. Z-scores corresponding to each cell-cycle stage were transformed into probabilities using the 'pnorm' function in R. Furthermore, the cell-cycle stage displaying the highest probability was designated as the most probable cell stage.

ACR identification

Upon segregating the comprehensive single-base resolution Tn5 insertion sites BED dataset into distinct subsets aligned with annotated cell types, we executed the MACS2 tool[60] (v2.2.7.1) for

precise peak identification per cell type. Notably we employed non-default parameters, specifically: –extsize 150, –shift -75, -nomodel -keep-dup all. To mitigate potential false positives, a permutation strategy was applied, generating an equal number of peaks based on regions that were mappable and non-exonic. This approach encompassed the assessment of Tn5 insertion sites and density within both the original and permuted peak groups. By scrutinizing the permutation outcomes, we devised empirically derived false discovery rate (FDR) thresholds specific to each cell type. This entailed determining the minimum Tn5 density score within the permutation cohort where the FDR remained < 0.05. To further eliminate peaks that exhibited significant overlap with nucleosomes, we applied the NucleoATAC tool[69] (v0.2.1) to forecast potential nucleosome placements. Peaks that featured over 50% alignment with predicted nucleosomes were systematically removed. The average fragment size of reads overlapping with the peaks were calculated and the peaks with the average fragment size > 150 bp were filtered out. Ultimately, the pool of peaks for each cell type was amalgamated and fine-tuned, yielding 500-base pair windows that were centered on the summit of ACR coverage.

Identification of cell-type-specific ACRs in rice atlas

To identify cell-type-specific ACRs in the rice atlas, we implemented a series of cutoffs to determine whether the peak is accessible for a specific cell type. For each cell type, we first normalized the read coverage depth obtained from the MACS2 tool divided by total count of reads, and ensured that the max of normalized coverage within the peak exceeded a threshold of 2. Additionally, we calculated Tn5 integration sites per peak, filtering out peaks with fewer than 20 integration sites. Subsequently, we constructed a peak by cell types matrix based on Tn5 integration sites. This matrix underwent normalization using the 'cpm' function wrapped in edgeR[70] (v3.38.1) and 'normalize.quantiles' function wrapped within preprocessCore[71] (v1.57.1) in the R programming environment. To further refine our selection, a threshold of 2 was set for the counts per million (cpm) value per peak per cell type. Peaks that satisfied these distinct cutoff criteria were deemed accessible in the designated cell types. We identified these peaks as displaying chromatin accessibility in anywhere from one to ten of the primary cell types among the total pool of 60 main cell types.

Identification of cell-type-specific ACRs in leaf tissue across five species

Cell-type-specific ACRs were identified using the same method as found in Mendieta et al. 2024[20]. In brief we utilized a modified entropy metric combined with a bootstrapping approach[72]. For each cell-type and species a sample of 250 cells were taken 5000 times with replacement and a specificity metric was calculated. This specificity metric was then compared against a series of 5000 null distributions consisting of a random shuffle of 250 cells which were of mixed cell populations. Then, a nonparametric test was used to the median real bootstrap specificity score versus the null, and ACRs were labeled as cell-type-specific if they had a p value less than 0.0001. If ACRs were found to be specific to more than one cell type and less than four, these were labeled as being cell-type-restricted.

Correlation between chromatin accessibility of TF genes and motif deviation

We sourced rice and *A. thaliana* TFs from PlantTFDB[73] (v4.0) database. To identify rice orthologs of *A. thaliana* TFs, we employed BLAST[59] (BLAST+; v2.11.0) by utilizing protein fasta alignments with an E-value threshold of 1e-5 used for significance. Alignments were restricted to fasta sequences categorized as TFs from either species. To further refine the putative

orthologs, we applied filters based on functional similarity to *A. thaliana* TFs. Alignments with less than 15% identity were excluded, along with rice TFs associated with distinct families. From the remaining candidates, we selected the orthologs demonstrating the highest Pearson correlation coefficient concerning the motif deviation scores. Motif deviation scores of specific TF motifs within nuclei were computed *via* chromVAR[74] (v1.18.0)

Linear-model based motif enrichment analysis

We employed the FIMO tool from the MEME suite[75] (v5.1.1) with a significance threshold of p value $< 10^{-5}$ to predict motif locations. The motif frequency matrix used was sourced from the JASPAR plants motif database[76] (v9). Subsequently, we constructed a binarized peak-by-motif matrix and a motif-by-cell count matrix. This involved multiplying the peak-by-cell matrix with the peak-by-motif matrix. To address potential overrepresentation and computational efficiency, downsampling was implemented. Specifically, we standardized the cell count by randomly selecting 412 cells per cell type per species. This count represents the lowest observed cell count for a given cell type across all species. For each cell type annotation, total motif counts were predicted through negative binomial regression. This involved two input variables: an indicator column for the annotation, serving as the primary variable of interest, and a covariate representing the logarithm of the total number of nonzero entries in the input peak matrix for each cell. The regression provided coefficients for the annotation indicator column and an intercept. These coefficients facilitated the estimation of fold changes in motif counts for the annotation of interest in relation to cells from all other annotations. This iterative process was conducted for all motifs across all cell types. The obtained p values were adjusted using the Benjamini-Hochberg procedure to account for multiple comparisons. Finally, enriched motifs were identified by applying a dual filter criterion: corrected p values < 0.01, fold-change of the top enriched TF motif in cell type-specific peaks for all cell types should be over 1, and beta (motif enrichment score) > 0.05 or beta > 0.

Binomial test-based motif enrichment analysis

To assess the enrichment of motifs in a target set of ACRs, we performed analysis for each specific motif. We randomly selected an equivalent number of ACRs as found in the target set, repeating this process 100 times. Notably, the randomly selected ACR set did not overlap with the actual target set of ACRs. Following this, we computed the average ratio of ACRs capturing the motif within the null distribution.

Subsequently, we executed an exact binomial test[77], wherein we set this ratio as the hypothesized probability of success. The number of ACRs overlapping the motif in the target set was considered the number of successes, while the total number of ACRs in the target set represented the number of trials. The alternative hypothesis was specified as 'greater.' This meticulous approach allowed us to robustly evaluate and identify significant motif enrichments within the target set of ACRs.

Construction of control sets for enrichment tests

To perform comparative analysis of expression levels and chromatin accessibility of genes surrounding broad ACRs under and outside of H3K27me3 peaks, we sampled the same number of ACRs per cell type regarding the broad ACRs not under H3K27me3 peaks. This step is to

make sure that their nearby gene chromatin accessibility exhibited similar values compared to the broad ACRs under the H3K27me3 peaks.

To check if the H3K27me3-broad-ACRs could significantly capture the known PREs and capture the EMF2b ChIP-seq peaks, we generated control sets by randomly selecting not-H3K27me3-broad-ACR instances 100 times, yielding a mean number value for the control sets. The binomial test *p* value was calculated by comparing the mean ratio to the observed number of H3K27me3-broad ACRs overlapped with the PREs.

To test if H3K27me3-broad ACRs in *O. sativa*, *Z. mays*, and *S. bicolor* significantly capture six known motifs, we generated control sets by simulating sequences with the same length as ACRs 100 times, yielding a mean proportion for the control sets. The binomial test p value was calculated by comparing the mean ratio to the observed overlapping ratio of H3K27me3-broad ACRs capturing the motifs.

Identification of syntenic regions

Identification of syntenic gene blocks was done using the GENESPACE[78] (v1.4). In brief, to establish orthologous relationships between ACR sequences, ACRs in the *O. sativa* genome were extended to incorporate the two closest gene models for a 'query block' since GENESPACE only draws relationships between protein coding sequences. Then the GENESPACE function 'query_hits' was used with the argument 'synOnly = TRUE' to retrieve syntenic blocks. The resulting syntenic hits were further filtered to allow only a one to one relationship between *O.sativa* and the corresponding species. The corresponding syntenic blocks were then named and numbered, and both the genes and genomic coordinates were recorded.

To further identify corresponding ACRs within these blocks we set up a BLASTN pipeline[79] (v2.13.0). For each comparison of species, using *O. sativa* as the reference the underlying nucleotide sequences of the syntenic regions were extracted using Seqkit, and used as the blast reference database[80] (v2.5.1). The sequences underlying the ACRs within the same syntenic region in a different species were then used as the query. The blast was done using the following parameters to allow for alignment of shorter sequences '-task blastn-short -evalue 1e-3 -max_target_seqs 4-word_size 7 -gapopen 5 -gapextend 2 -penalty -1 -reward 1 -outfmt 6'. This procedure was run for each syntenic region separately for all species comparisons. The resulting BLASTN files were combined, and then filtered using a custom script. Alignments were only considered valid if the e-value passed a stringent threshold of 1e-3, and the alignment was greater than 20 nucleotides. The resulting filtered BLAST files, and the BED files generated from these BLAST files allowed us to draw our relationships between ACRs in the corresponding syntenic space.

Estimation of conservation scores

Conservation scores were predicted using PhyloP[81] (v1.0), where values are scaled between 0 to 1, with one being highly conserved and 0 being non-conserved. Phylogenies to train PhyloP were

generated using PhyloFit[82] (v1.0), and neutral and conserved sequences were identified using the whole genome aligner progressive cactus.

ChIP-seq analysis

The clean reads of EMF2b were downloaded from Tan's study[44]. The reads were mapped to the rice reference genome[50] (v7.0) using bowtie2[83] (v2.5.2) with the following parameters: '--very-sensitive --end-to-end'. Reads with MAPQ > 5 were used for thesubsequent analysis. Aligned reads were sorted and duplicated reads were removed using SAMtools[56] (v1.7). Peak calling was performed using epic2[84] with the following parameters: '-fdr 0.01 --bin-size 150 --gaps-allowed 1'. The peak 'BED' and 'BIGWIG' files of H3K27me3 ChIP-seq data for leaf, root, and panicle rice organs were downloaded from RiceENCODE[85] (http://glab.hzau.edu.cn/RiceENCODE/)

Additional resources

Cell-type resolved data can be viewed through our public Plant Epigenome JBrowse Genome Browser[86] (http://epigenome.genetics.uga.edu/PlantEpigenome/index.html).

Data Availability

scATAC-seq data encompassing 18 libraries from nine organs were accessible in NCBI (PRJNA1007577/GSE252040; https://dataview.ncbi.nlm.nih.gov/object/PRJNA1007577?reviewer=kgarq48dii11vomg44kgr1jq66; PRJNA1052039; https://dataview.ncbi.nlm.nih.gov/object/PRJNA1052039?reviewer=flhu9sl84o5m999r1ph8tlmmbg).

Acknowledgements

This research was funded by the National Science Foundation (IOS-2134912) to SRW and RJS. APM and JPM were supported by the National Institutes of Health (K99GM144742) and (T32GM142623), respectively.

Contributions

R.J.S., H.Y., J.P.M., and S.R.W. designed and conceived experiments and managed the project. X.Z., Y.L., Z.L., X.T., S.Z., Y.W., and H.Y. participated in material collection and sample processing. H.Y., J.P.M., and T.R. performed the bioinformatics analyses. H.Y. and J.P.M. wrote the manuscript. Y.L. Y.W. and Z.L. contributed to marker validation. R.J.S., J.P.M., A.P.M., D.W., S.Z., and S.R.W. edited the manuscript.

References cited

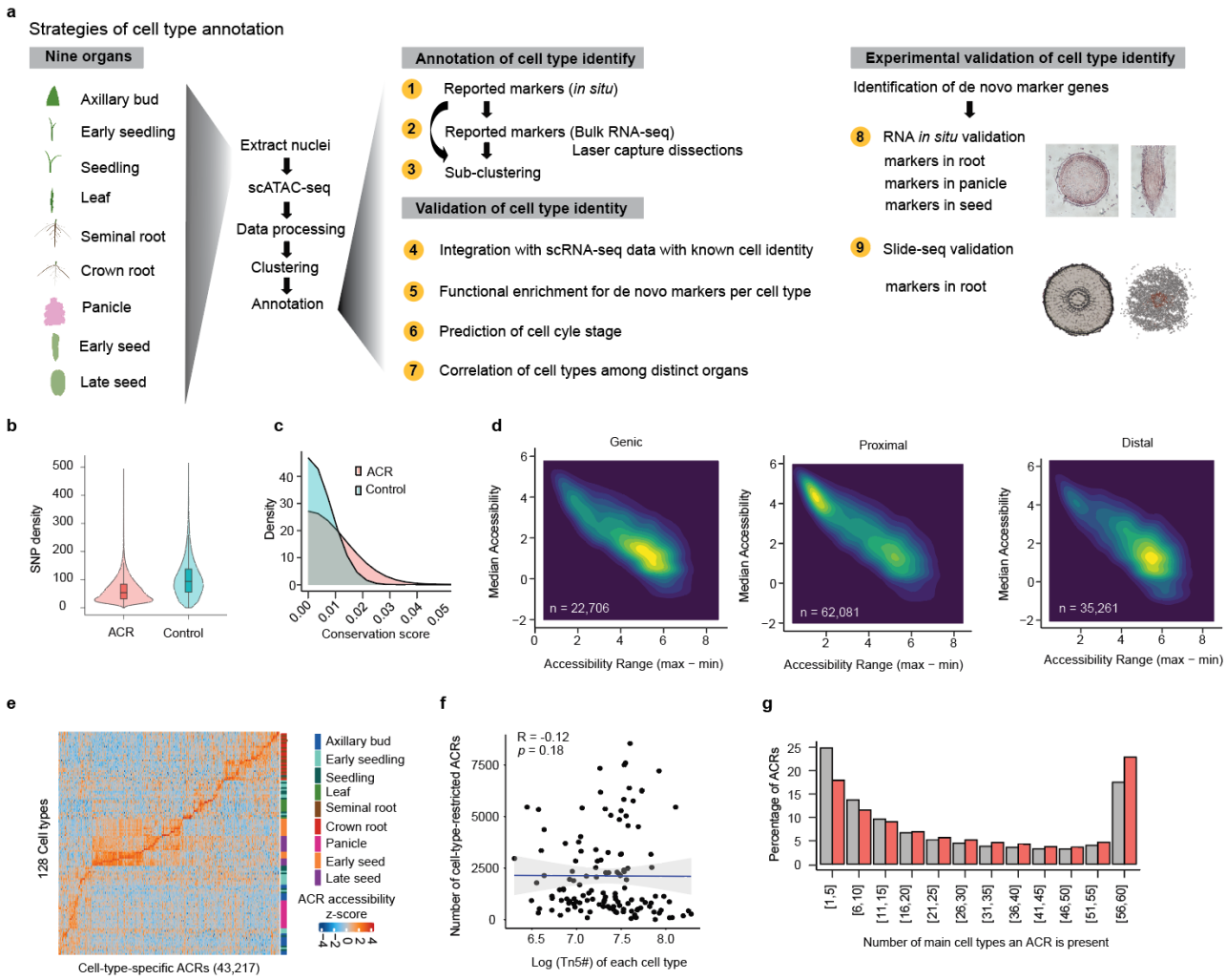
1. Preissl, S., Gaulton, K. J. & Ren, B. Characterizing cis-regulatory elements using single-cell epigenomics. *Nature Reviews Genetics* **24**, 21-43 (2023).
2. Wittkopp, P. J. & Kalay, G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics* **13**, 59-69 (2012).
3. Ricci, W. A. *et al.* Widespread long-range cis-regulatory elements in the maize genome. *Nature plants* **5**, 1237-1249 (2019).
4. Oka, R. *et al.* Genome-wide mapping of transcriptional enhancer candidates using DNA and chromatin features in maize. *Genome biology* **18**, 1-24 (2017).
5. Maher, K. A. *et al.* Profiling of accessible chromatin regions across multiple plant species and cell types reveals common gene regulatory principles and new control modules. *The Plant Cell* **30**, 15-36 (2018).
6. Lu, Z. *et al.* The prevalence, evolution and chromatin signatures of plant regulatory elements. *Nature Plants* **5**, 1250-1259 (2019).
7. Reynoso, M. A. *et al.* Evolutionary flexibility in flooding response circuitry in angiosperms. *Science* **365**, 1291-1295 (2019).
8. Kajala, K. *et al.* Innovation, conservation, and repurposing of gene function in root cell type development. *Cell* **184**, 3333-3348. e3319 (2021).
9. Dorrity, M. W. *et al.* The regulatory landscape of Arabidopsis thaliana roots at single-cell resolution. *Nature communications* **12**, 3334 (2021).
10. Cusanovich, D. A. *et al.* A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell* **174**, 1309-1324. e1318 (2018).
11. Domcke, S. *et al.* A human cell atlas of fetal chromatin accessibility. *Science* **370**, eaba7612 (2020).
12. Lu, Z. *et al.* Tracking cell-type-specific temporal dynamics in human and mouse brains. *Cell* **186**, 4345-4364. e4324 (2023).
13. Farmer, A., Thibivilliers, S., Ryu, K. H., Schiefelbein, J. & Libault, M. Single-nucleus RNA and ATAC sequencing reveals the impact of chromatin accessibility on gene expression in Arabidopsis roots at the single-cell level. *Molecular Plant* **14**, 372-383 (2021).
14. Tu, X., Marand, A. P., Schmitz, R. J. & Zhong, S. A combinatorial indexing strategy for low-cost epigenomic profiling of plant single cells. *Plant Communications* **3** (2022).
15. Nobori, T. *et al.* Time-resolved single-cell and spatial gene regulatory atlas of plants under pathogen attack. *bioRxiv*, 2023.2004. 2010.536170 (2023).
16. Marand, A. P., Chen, Z., Gallavotti, A. & Schmitz, R. J. A cis-regulatory atlas in maize at single-cell resolution. *Cell* **184**, 3041-3055. e3021 (2021).
17. Feng, D. *et al.* Chromatin accessibility illuminates single-cell regulatory dynamics of rice root tips. *BMC biology* **20**, 274 (2022).
18. Zhang, L. *et al.* Asymmetric gene expression and cell-type-specific regulatory networks in the root of bread wheat revealed by single-cell multiomics analysis. *Genome Biology* **24**, 65 (2023).
19. Swift, J. *et al.* Single nuclei sequencing reveals C4 photosynthesis is based on rewiring of ancestral cell identity networks. *bioRxiv*, 2023.2010. 2026.562893 (2023).
20. Mendieta, J. P. *et al.* Investigating the cis-Regulatory Basis of C3 and C4 Photosynthesis in Grasses at Single-Cell Resolution. *bioRxiv*, 2024.2001. 2005.574340 (2024).
21. Lv, Z., Zhao, W., Kong, S., Li, L. & Lin, S. Overview of molecular mechanisms of plant leaf development: a systematic review. *Frontiers in Plant Science* **14** (2023).
22. Le Hir, R. & Bellini, C. The plant-specific Dof transcription factors family: new players involved in vascular system development and functioning in Arabidopsis. *Frontiers in Plant Science* **4**, 164 (2013).

23 Dai, X. *et al.* Chromatin and regulatory differentiation between bundle sheath and mesophyll cells in maize. *The Plant Journal* **109**, 675-692 (2022).
24 Yanagisawa, S. Dof1 and Dof2 transcription factors are associated with expression of multiple genes involved in carbon metabolism in maize. *The Plant Journal* **21**, 281-288 (2000).
25 Ueno, O., Kawano, Y., Wakayama, M. & Takeda, T. Leaf vascular systems in C3 and C4 grasses: a two-dimensional analysis. *Annals of Botany* **97**, 611-621 (2006).
26 Gao, Z., Shen, W. & Chen, G. Uncovering C4-like photosynthesis in C3 vascular cells. *Journal of Experimental Botany* **69**, 3531-3540 (2018).
27 Ingram, P., Dettmer, J., Helariutta, Y. & Malamy, J. E. Arabidopsis Lateral Root Development 3 is essential for early phloem development and function, and hence for normal root system development. *The Plant Journal* **68**, 455-467 (2011).
28 Chen, X. *et al.* SQUAMOSA promoter-binding protein-like transcription factors: Star players for plant growth and development. *Journal of integrative plant biology* **52**, 946-951 (2010).
29 Denyer, T. *et al.* Spatiotemporal developmental trajectories in the Arabidopsis root revealed using high-throughput single-cell RNA sequencing. *Developmental cell* **48**, 840-852. e845 (2019).
30 Fang, J. *et al.* The URL1–ROC5–TPL2 transcriptional repressor complex represses the ACL1 gene to modulate leaf rolling in rice. *Plant Physiology* **185**, 1722-1744 (2021).
31 Horstman, A. *et al.* AIL and HDG proteins act antagonistically to control cell proliferation. *Development* **142**, 454-464 (2015).
32 Yocca, A. E., Lu, Z., Schmitz, R. J., Freeling, M. & Edger, P. P. Evolution of conserved noncoding sequences in Arabidopsis thaliana. *Molecular Biology and Evolution* **38**, 2692-2703 (2021).
33 Hendelman, A. *et al.* Conserved pleiotropy of an ancient plant homeobox gene uncovered by cis-regulatory dissection. *Cell* **184**, 1724-1739. e1716 (2021).
34 de Velde, J. V., Heyndrickx, K. S. & Vandepoele, K. Inference of transcriptional networks in Arabidopsis through conserved noncoding sequence analysis. *The Plant Cell* **26**, 2729-2745 (2014).
35 Woolfe, A. *et al.* Highly conserved non-coding sequences are associated with vertebrate development. *PLoS biology* **3**, e7 (2005).
36 Babarinde, I. A. & Saitou, N. Genomic locations of conserved noncoding sequences and their proximal protein-coding genes in mammalian expression dynamics. *Molecular biology and evolution* **33**, 1807-1817 (2016).
37 Song, B. *et al.* Conserved noncoding sequences provide insights into regulatory sequence and loss of gene expression in maize. *Genome research* **31**, 1245-1257 (2021).
38 Cao, R. *et al.* Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science* **298**, 1039-1043 (2002).
39 Xiao, J. *et al.* Cis and trans determinants of epigenetic silencing by Polycomb repressive complex 2 in Arabidopsis. *Nature genetics* **49**, 1546-1552 (2017).
40 Schmitz, R. J., Grotewold, E. & Stam, M. Cis-regulatory sequences in plants: Their importance, discovery, and future challenges. *The plant cell* **34**, 718-741 (2022).
41 Ouyang, W. *et al.* Haplotype mapping of H3K27me3-associated chromatin interactions defines topological regulation of gene silencing in rice. *Cell Reports* **42** (2023).
42 Yu, Y., Zhang, H., Long, Y., Shu, Y. & Zhai, J. Plant public RNA-seq database: a comprehensive online database for expression analysis of~ 45 000 plant public RNA-seq libraries. *Plant Biotechnology Journal* **20**, 806 (2022).
43 Tonosaki, K. & Kinoshita, T. Possible roles for polycomb repressive complex 2 in cereal endosperm. *Frontiers in plant science* **6**, 144 (2015).

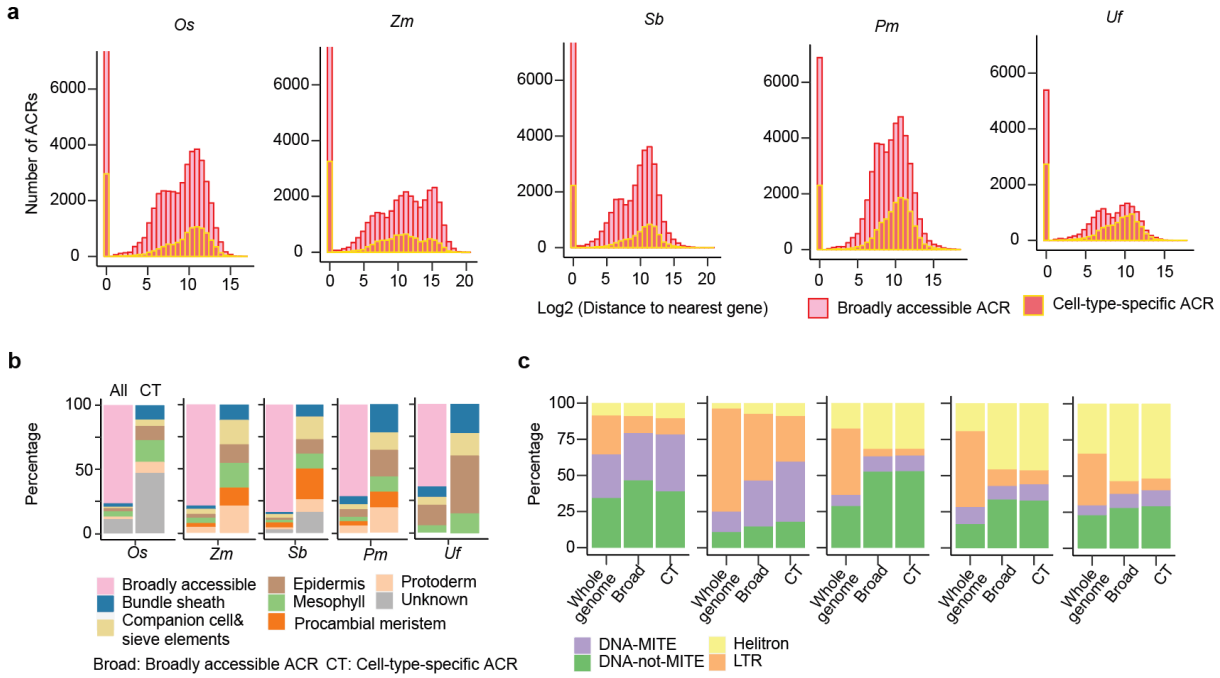
44 Tan, F.-Q. *et al.* A coiled-coil protein associates Polycomb Repressive Complex 2 with KNOX/BELL transcription factors to maintain silencing of cell differentiation-promoting genes in the shoot apex. *The Plant Cell* **34**, 2969-2988 (2022).
45 Bai, X. *et al.* Duplication of an upstream silencer of FZP increases grain yield in rice. *Nature Plants* **3**, 885-893 (2017).
46 Javelle, M., Vernoud, V., Rogowsky, P. M. & Ingram, G. C. Epidermis: the formation and functions of a fundamental plant tissue. *New Phytologist* **189**, 17-39 (2011).
47 Ciren, D., Zebell, S. & Lippman, Z. Extreme restructuring of cis-regulatory regions controlling a deeply conserved plant stem cell regulator. *bioRxiv*, 2023.2012. 2020.572550 (2023).
48 Zhang, X., Marand, A. P., Yan, H. & Schmitz, R. J. Massive-scale single-cell chromatin accessibility sequencing using combinatorial fluidic indexing. *BioRxiv* (2023).
49 Zheng, G. X. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature communications* **8**, 14049 (2017).
50 Ouyang, S. *et al.* The TIGR rice genome annotation resource: improvements and new features. *Nucleic acids research* **35**, D883-D887 (2007).
51 Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology* **33**, 495-502 (2015).
52 Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell systems* **8**, 281-291. e289 (2019).
53 Rodriques, S. G. *et al.* Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463-1467 (2019).
54 Stickels, R. R. *et al.* Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nature biotechnology* **39**, 313-319 (2021).
55 Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nature biotechnology* **37**, 925-936 (2019).
56 Li, H. *et al.* The sequence alignment/map format and SAMtools. *bioinformatics* **25**, 2078-2079 (2009).
57 Institute, B. Picard toolkit. *Broad Institute, GitHub repository* (2019).
58 Smit, A. F. Repeat-Masker Open-3.0. *http://www. repeatmasker. org* (2004).
59 Camacho, C. *et al.* BLAST+: architecture and applications. *BMC bioinformatics* **10**, 1-9 (2009).
60 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**, 1-9 (2008).
61 DeBruine, Z. J., Melcher, K. & Triche Jr, T. J. Fast and robust non-negative matrix factorization for single-cell experiments. *bioRxiv*, 2021.2009. 2001.458620 (2021).
62 McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
63 Hua, L. *et al.* The bundle sheath of rice is conditioned to play an active role in water transport as well as sulfur assimilation and jasmonic acid synthesis. *The Plant Journal* **107**, 268-286 (2021).
64 Itoh, J.-I. *et al.* Genome-wide analysis of spatiotemporal gene expression patterns during early embryogenesis in rice. *Development* **143**, 1217-1227 (2016).
65 Wu, T.-Y., Müller, M., Gruissem, W. & Bhullar, N. K. Genome wide analysis of the transcriptional profiles in different regions of the developing rice grains. *Rice* **13**, 1-19 (2020).
66 Van Dijk, D. *et al.* Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716-729. e727 (2018).
67 Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature methods* **16**, 1289-1296 (2019).

68 Pettkó-Szandtner, A. *et al.* Core cell cycle regulatory genes in rice and their expression profiles across the growth zone of the leaf. *Journal of plant research* **128**, 953-974 (2015).
69 Schep, A. N. *et al.* Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome research* **25**, 1757-1770 (2015).
70 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 (2010).
71 Bolstad, B. M. & Bolstad, M. B. M. Package 'preprocessCore'. (2013).
72 Zhang, T.-Q., Chen, Y., Liu, Y., Lin, W.-H. & Wang, J.-W. Single-cell transcriptome atlas and chromatin accessibility landscape reveal differentiation trajectories in the rice root. *Nature communications* **12**, 2053 (2021).
73 Jin, J. *et al.* PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic acids research*, gkw982 (2016).
74 Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nature methods* **14**, 975-978 (2017).
75 Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017-1018 (2011).
76 Castro-Mondragon, J. A. *et al.* JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic acids research* **50**, D165-D173 (2022).
77 Wagner-Menghin, M. M. Binomial test. *Encyclopedia of statistics in behavioral science* (2005).
78 Lovell, J. T. *et al.* GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *Elife* **11**, e78526 (2022).
79 Sayers, E. W. *et al.* Database resources of the national center for biotechnology information. *Nucleic acids research* **49**, D10 (2021).
80 Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PloS one* **11**, e0163962 (2016).
81 Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research* **20**, 110-121 (2010).
82 Siepel, A. & Haussler, D. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Molecular biology and evolution* **21**, 468-488 (2004).
83 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357-359 (2012).
84 Stovner, E. B. & Sætrom, P. epic2 efficiently finds diffuse domains in ChIP-seq data. *Bioinformatics* **35**, 4392-4393 (2019).
85 Xie, L. *et al.* RiceENCODE: A comprehensive epigenomic database as a rice Encyclopedia of DNA Elements. *Molecular Plant* **14**, 1604-1606 (2021).
86 Hofmeister, B. T. & Schmitz, R. J. Enhanced JBrowse plugins for epigenomics data visualization. *BMC bioinformatics* **19**, 1-6 (2018).
87 Shimadzu, S., Furuya, T. & Kondo, Y. Molecular mechanisms underlying the establishment and maintenance of vascular stem cells in Arabidopsis thaliana. *Plant and Cell Physiology* **64**, 274-283 (2023).
88 Otero, S. & Helariutta, Y. Companion cells: a diamond in the rough. *Journal of experimental botany*, erw392 (2016).
89 Ramachandran, V. *et al.* Plant-specific Dof transcription factors VASCULAR-RELATED DOF1 and VASCULAR-RELATED DOF2 regulate vascular cell differentiation and lignin biosynthesis in Arabidopsis. *Plant molecular biology* **104**, 263-281 (2020).

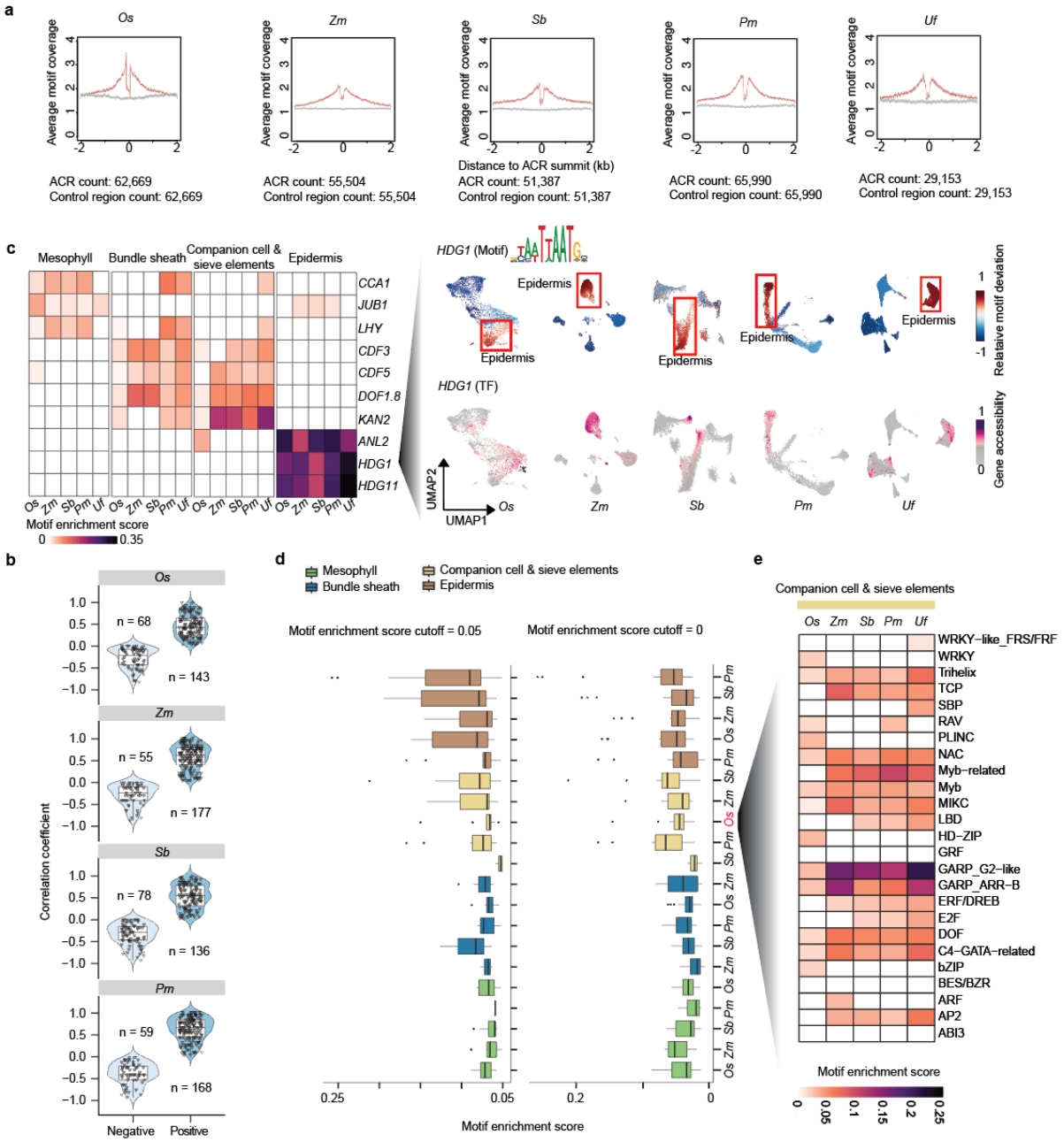
90 Emery, J. F. *et al.* Radial patterning of Arabidopsis shoots by class III HD-ZIP and KANADI genes. *Current Biology* **13**, 1768-1774 (2003).
91 Kubo, H., Kishi, M. & Goto, K. Expression analysis of ANTHOCYANINLESS2 gene in Arabidopsis. *Plant science* **175**, 853-857 (2008).
92 Wu, R. *et al.* CFL1, a WW domain protein, regulates cuticle development by modulating the function of HDG1, a class IV homeodomain transcription factor, in rice and Arabidopsis. *The Plant Cell* **23**, 3392-3411 (2011).
93 Amanda, D. *et al.* DEFECTIVE KERNEL1 (DEK1) regulates cell walls in the leaf epidermis. *Plant Physiology* **172**, 2204-2218 (2016).



Extended Data Fig. 1. Pipeline of cell identity annotation and quality control of rice ACRs. a, A pipeline of cell identity annotation corresponding to nine strategies shown in Supplementary Note 1. **b-c,** A comparison between ACRs (n = 120,048) and control regions (n = 120,048) in terms of SNP density (**b**) and conservation scores (**c**). The ACRs exhibited lower SNP density and higher conservation scores compared to the control, suggesting functional importance. The control set was created to match each ACR by excluding exons, ACRs, and unmappable regions, and they possess a similar GC content (< 5% average difference) compared to the positive set. **d,** A two-dimensional density plot illustrates the median chromatin accessibility across 128 cell types for 120,048 ACRs, along with the range of chromatin accessibility (calculated as the difference between the maximum and minimum values). **e,** Row normalized chromatin accessibility z-scores for ACRs across cell types. **f,** Spearman correlation is shown between the total number of Tn5 insertions and number of cell-type-restricted ACRs across all cell types. The identification of cell-type-specific ACRs was independent of sequencing depth, as evidenced by the lack of correlation between Tn5 insertions and the number of cell-type-specific ACRs per cell type. **g,** Percentage of ACRs within main cell types in rice, distributed across syntenic and non-syntenic regions shared with four other grass species (*Z. mays*, *S. bicolor*, *P. miliaceum*, and *U. fusca*).

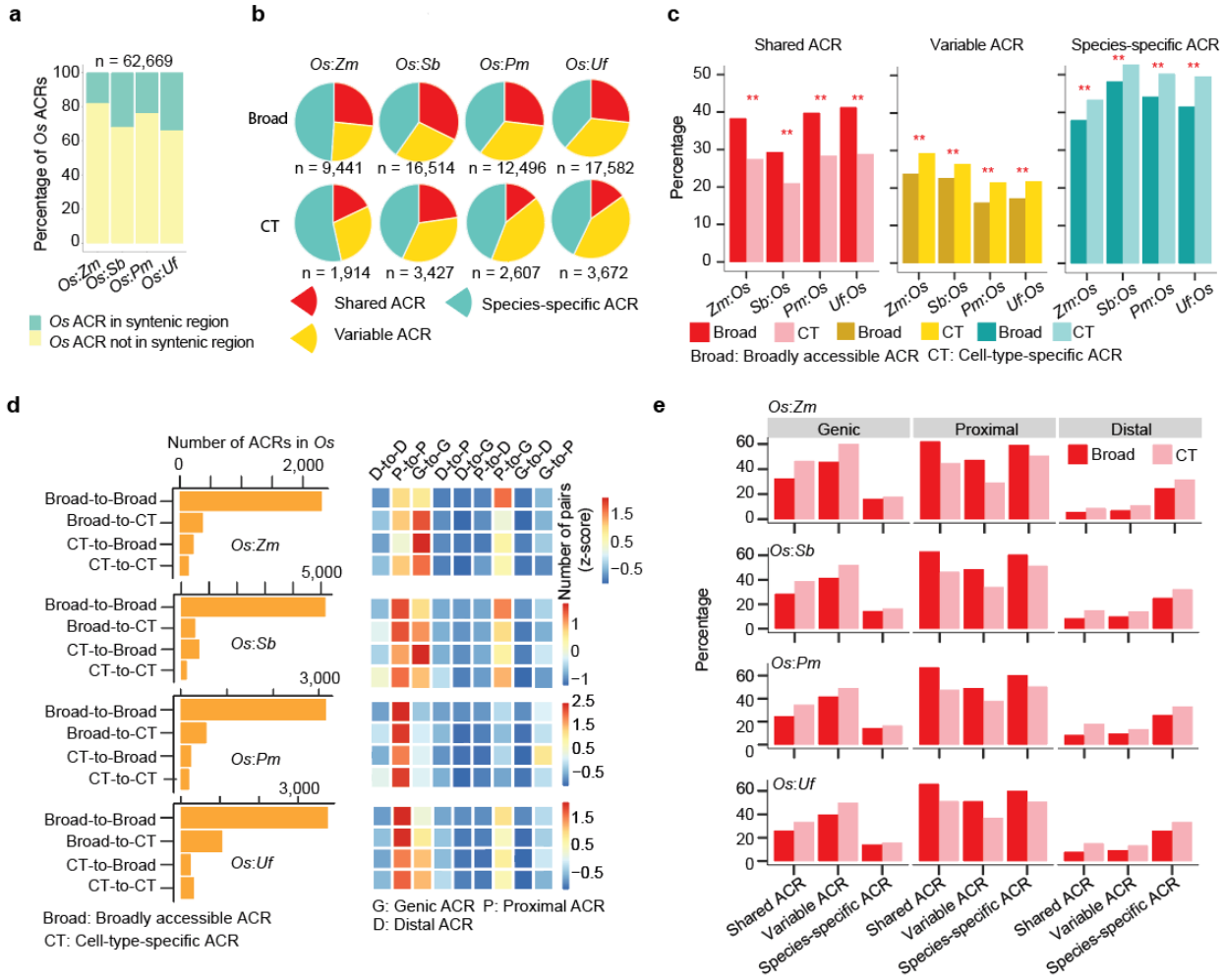


Extended Data Fig. 2. Characterization of ACRs across five grass species. a, Distribution of distance of ACRs to their nearest genes in broadly accessible and cell-type-specific ACRs across five species. **b,** The distribution of cell-type-specific ACRs across different cell types was characterized by a similar percentage representation. **c,** The compositions of the family of transposable elements (TEs) overlapping with broad and cell-type-specific ACRs illustrate comparable compositions across TE super families, distinct from the whole genome.

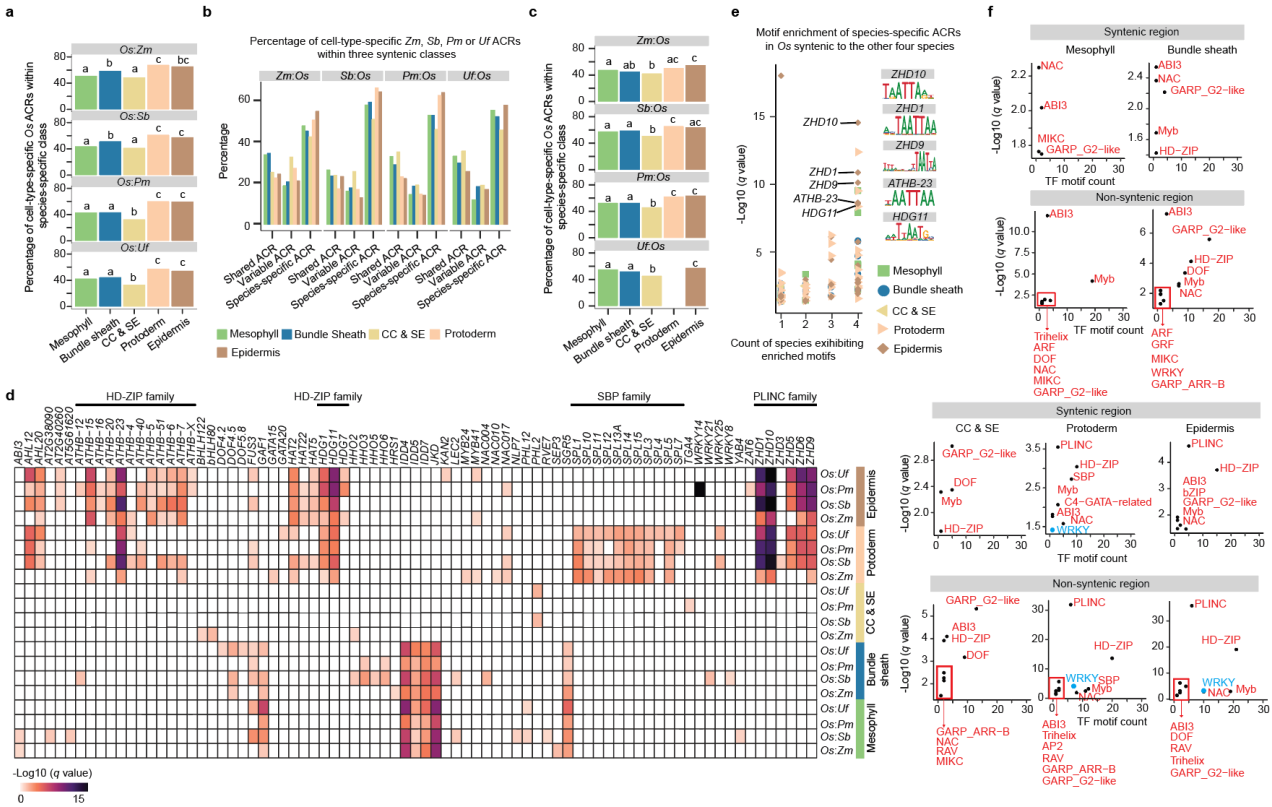


Extended Data Fig. 3. Examination of the chromatin accessibility of motifs enriched on specific cell types across species. a, Distribution of distance of ACRs to their nearest genes in broadly accessible and cell-type-specific ACRs across five species. **b,** Mean motif coverage centered on ACRs and control regions for each species, with 95% confidence intervals shown as a shaded polygon. **c,** Spearman correlation between chromatin accessibility of TF genes and motif deviation. Note that for *Uf* species, the analysis was omitted due to the limited number of cell types (four) available for spearman correlation, making it statistically unfeasible. **d**, Deviations of ten motifs displaying enrichment consistent with the gene body chromatin accessibility of their associated TFs. The left heatmap panel illustrates ten TF motif enrichments, consistent with the known TF dynamics among cell types [*CIRCADIAN CLOCK ASSOCIATED 1* (*CCA1*), LATE ELONGATED HYPOCOTYL (*LHY*) and *JUNGBRUNNEN1* (*JUB1*)[87]: mesophyll; *CYCLING DOF FACTOR 3* (*CDF3*) and *CDF5*[88]: companion cells; *DOF1.8*[89] and *KANADI2* (*KAN2*)[90]: vascular-related cells; *ANTHOCYANINLESS2* (*ANL2*)[91], *HOMEODOMAIN GLABROUS*

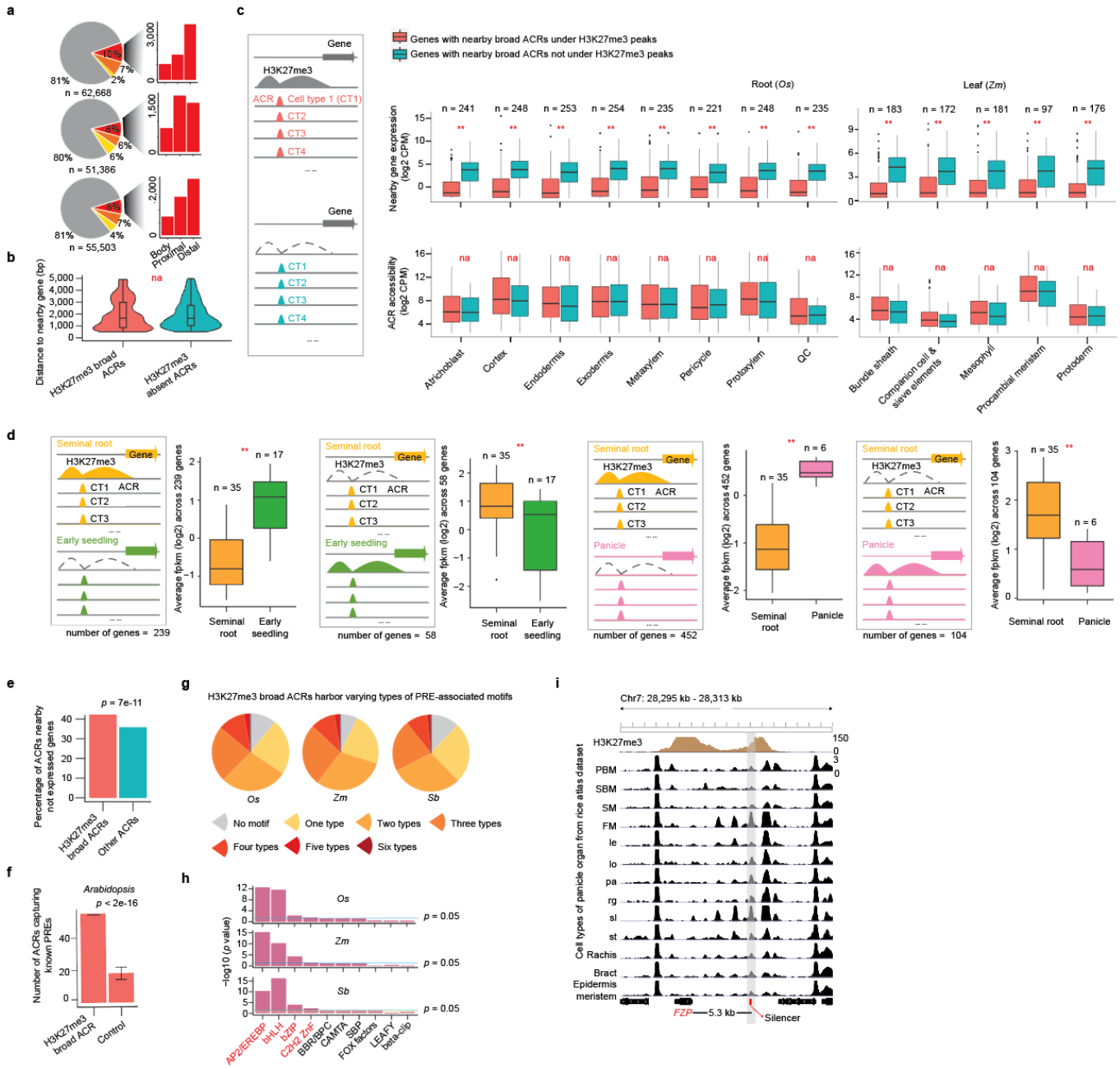
(*HDG1*)[92], and *HDG11*[93]]. The right panels highlight one motif (*HDG1*) enriched with ACRs in epidermis, where their cognate TFs are known to accumulate. **e,** Presenting box (left) and heatmap (right) plots illustrating the collapsed motif enrichment patterns into TF motif families across various species for each cell type. Each dot within the box represents a TF motif family. The reduced motif enrichment cutoff setting reveals additional motif families enriched in companion cells and sieve elements, compared to the higher beta cutoff setting as shown in Fig. 2d. It's worth noting that we still observed lower enrichment scores in *O. sativa* compared to other species.



Extended Data Fig. 4. Characterization of ACRs within syntenic regions. a, The bar plot illustrates the percentage of *O. sativa* ACRs within syntenic and non-syntenic regions when compared to another species. **b,** The pie charts depict percentages of broadly accessible (Broad) and cell-type-specific (CT) ACRs within three classes of ACR conservation in Fig. 3b. The total number of ACR in *O. sativa* was written below the pie charts. **c,** The percentage of broad and cell-type-specific ACRs using *Z. mays*, *S. bicolor*, *P. miliaceum*, or *U. fusca* as the baseline underlying three classes shown in Fig. 3b. The significance test was done by using the Fisher's exact test (alternative = 'greater'). '**' denotes a p value < 0.01, and '*' denotes a p value < 0.05. **d,** Characterization of ACRs under 'shared ACR' class and their proximity categorizations. The left panel illustrates the number of ACRs under the 'shared ACR' class, categorized into four combinations based on broadly accessible and cell-type-specific ACRs. The right panel presents a heatmap of ACR categories based on their proximity to genes: genic ACRs overlapping genes, proximal ACRs (within 2 kb of genes), and distal ACRs (more than 2 kb away from the TSS). **e,** The percentage of shared, variable, and species-specific ACRs in genic, proximal, and distal manners for each species pair. The percentage for each class within genic, proximal, and distal groups collectively sum to 100%.

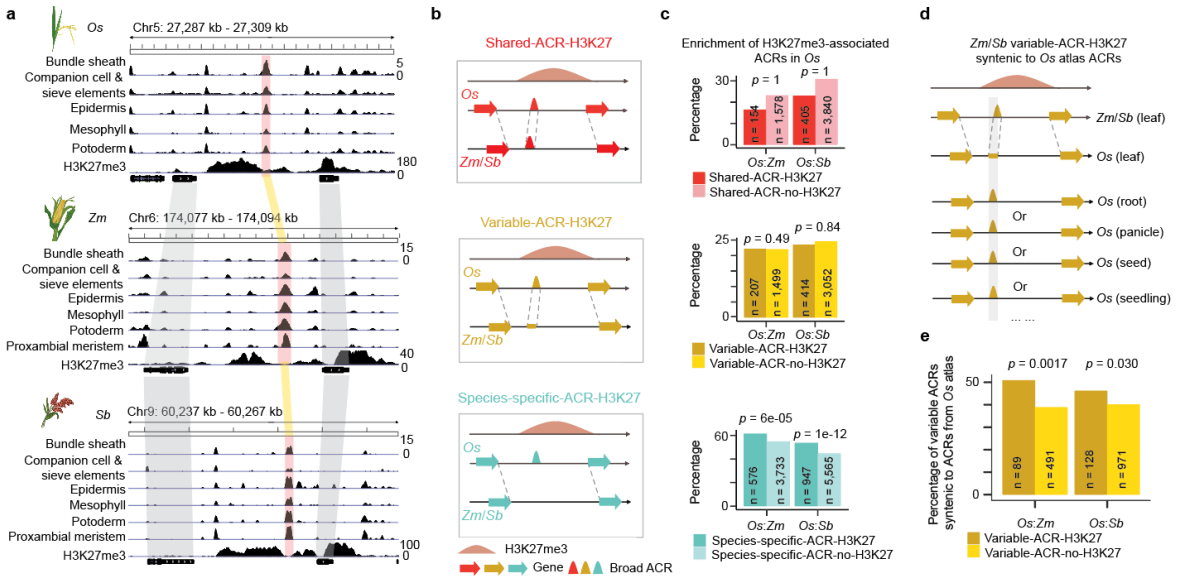


Extended Data Fig. 5. Epidermal cell specific ACRs are less conserved in sequence than other examined cell types. a, The analysis demonstrates an enrichment of cell-type-specific ACRs based on Fisher's exact test. This test assesses whether cell-type-specific ACRs are more likely to be situated in the 'species-specific' class, compared to other cell types. Statistically significant differences ($p < 0.05$) in all pairwise comparisons are denoted by distinct letters, determined using Fisher's exact test with the alternative set to 'greater'. Bars sharing the same letter indicate that they are not significantly ($p > 0.05$) different from each other. **b,** The proportion of cell-type-specific ACRs identified collectively across all cell types within each species pair under three syntenic classes. The percentage for each cell type within the three classes collectively sum to 100%. **c,** This panel has the same meaning as panel **a**, and corresponds to an enrichment test on panel **b**. *U. fusca*:*O. sativa* did not include a bar for protoderm as this cell type has not been identified in *U. fusca* scATAC-seq dataset. **d-e,** TF motif enrichment tests were performed in species-specific ACRs per species pair using *O. sativa* as the baseline (See Methods: Binomial test based motif enrichment analysis). The panel **d** presents a heat map illustrating motif enrichment per TF motif, while the panel **e** summarizes the count of species pairs showing enriched TF motifs. The top five motifs were highlighted with the lowest enriched *q* values for ACRs specific to epidermis cells. **f,** This panel utilized the same method as panel **d** to conduct TF motif enrichment tests in syntenic and non-syntenic ACRs in the *O. sativa*:*Z. mays* species pair (The results from other species pairs were shown in Supplementary Table 8). The '-Log10 (*q* value)' of each TF family was calculated by averaging '-Log10 (*q* value)' across all individual TF motifs within this family. TF family names were marked besides each dark dot, while the WRKY family was highlighted using blue color. The x-axis indicates the number of TF motifs detected to be enriched in the relative TF family.



Extended Data Fig. 6. Characterization of H3K27me3-broad ACRs potentially enriched for silencer CREs. a, Pie charts showcase the composition of different categories of ACRs, and bar plots display the number of ACRs within the H3K27me3-broad-ACR group. **b,** No significant difference in the distance to nearby genes was identified between H3K27me3-broad ACRs to H3K27me3-absent ACRs as shown in Fig. 5c. **c,** Comparative analysis of expression levels and chromatin accessibility of genes surrounding broad ACRs under and outside of H3K27me3 peaks. ** indicate *p* value < 0.01. **d,** Four categories of gene sets overlapped with H3K27me3 peaks and comparisons of their expression between bulk RNA-seq data libraries from two different organs. The 'n' above the box indicates the number of libraries used in this comparison. **e**, Percentage of H3K27me3-broad ACRs nearby genes exhibiting no expression across any cell type was significantly higher than the other ACRs associated with the not expressed genes. **f,** The

number of H3K27me3-broad-ACR capturing known 53 PREs is determined based on *A. thaliana* scATAC-seq data. The significance test was done by using the Fisher's exact test (alternative = 'greater'). **g,** Pie charts illustrate the comparison of different types of PRE-associated motifs harbored in H3K27me3-broad ACRs. **h**, Four TF families, highlighted in red, were significantly enriched in H3K27me3-broad ACRs. The motif data were collected from 568 TFs from *A. thaliana* belonging to 24 families within the JASPAR database[76]. The *p* value was computed using a hypergeometric test. **i,** A screenshot illustrates a H3K27me3-broad ACR harboring a reported silencer within an H3K27me3 peak in the panicle organ, located approximately 5.3 kb upstream of *FZP* gene. SBM: Secondary Branch Meristem. PBM: Primary Branch Meristem. SM: Spikelet Meristem. FM: Floral Meristem. le: Lemma. lo: Lodicule. pa: Palea. rg: Rudimentary glume. sl: Sterile lemma.



Extended Data Fig. 7. Evolutionary dynamics of H3K27me3-broad ACRs across different species. **a,** An example of a syntenic region containing H3K27me3-broad ACRs that were conserved across *O. sativa*, *Z. mays*, and *S. bicolor*. **b,** Three categories of *O. sativa* H3K27me3-broad ACRs that are syntenic to regions from another species. **c,** Comparison of ACR percentage underlying three classes corresponding to panel G associated with H3K27me3 or not. The *p* values displayed in the bottom bar plot were computed using Fisher's exact test. **d,** A sketch graph illustrates whether variable H3K27me3-broad ACRs in *Z. mays* and *S. bicolor* could capture ACRs derived from *O. sativa* atlas in Fig.1. **e,** The bar plot depicts the percentage of variable ACRs capturing the *O. sativa* atlas ACRs. The significance test was done by using the Fisher's exact test (alternative = 'greater').