nature plants

ARTICLES

https://doi.org/10.1038/s41477-019-0548-z

# The prevalence, evolution and chromatin signatures of plant regulatory elements

Zefu Lu[1], Alexandre P. Marand[1], William A. Ricci[2], Christina L. Ethridge[1], Xiaoyu Zhang[2*] and Robert J. Schmitz[1*]

**Chromatin accessibility and modification is a hallmark of regulatory DNA, the study of which led to the discovery of *cis*-regulatory elements (CREs). Here, we characterize chromatin accessibility, histone modifications and sequence conservation in 13 plant species. We identified thousands of putative CREs and revealed that distal CREs are prevalent in plants, especially in species with large and complex genomes. The majority of distal CREs have been moved away from their target genes by transposable-element (TE) proliferation, but a substantial number of distal CREs also seem to be created by TEs. Finally, plant distal CREs are associated with three major types of chromatin signatures that are distinct from metazoans. Taken together, these results suggest that CREs are prevalent in plants, highly dynamic during evolution and function through distinct chromatin pathways to regulate gene expression.**

Non-coding CREs, which exhibit high levels of chromatin accessibility, provide essential information to specify the level as well as the temporal and spatial patterns of gene expression[1–4]. In mammalian genomes, a large number of distal CREs have been found to possess specific chromatin features and dynamically interact with their target genes through long-distance chromatin looping[5–11]. The lack of distal CREs in the unicellular *Capsaspora owczarzaki*, an amoeba species that is sister to other animals, has led to the hypothesis that distal CREs are metazoan innovations that enable complex gene regulation that is critical for multicellularity[12]. However, the identification of a handful of distal CREs in plant genomes brings this notion into question[13,14]. Sequence variations at several distal CREs cause phenotypic variations that are important for the domestication of *Zea mays*[15–18], and variation in time to flowering in *Arabidopsis thaliana*[19–22] implied the existence of distal CREs in plants. Moreover, in *Z. mays*, single-nucleotide polymorphisms (SNPs) in accessible chromatin explain approximately 40% of the heritable variance of quantitative traits[23]. Although the genetic identification and functional validation of these distal CREs prove their existence in plant genomes, and accessible chromatin regions (ACRs) have been identified several plant species[24–30], the prevalence, chromatin signatures and mode of action of distal CREs remain unknown across most angiosperm genomes[31–33]. To address these questions, we performed a set of comparative genomic and epigenomic analyses in 13 angiosperm species. The results not only identified tens of thousands of putative plant CREs, including those located in promoters, terminators and distal to genes, but also showed that putative distal CREs are characterized by distinct chromatin features that are predictive of their mode of action on target genes.

## Results

**Distal accessible chromatin correlates with genome size.** Interactions between transcription factors (TFs) and DNA at CREs disrupts nucleosome formation, resulting in local ACRs[23,32,34–39]. The assay for transposase-accessible chromatin using sequencing (ATAC-seq) method enables genome-wide identification of ACRs that are probably putative CREs[37,39–42]. We therefore performed ATAC-seq using leaf tissues from 13 angiosperm species, including monocots and dicots with diverse genome sizes (~100–5,000 Mb), genome structure and gene content (18505–56044; Fig. 1a,b, Supplementary Fig. 1a). Accessible chromatin was found at many of the rare, albeit well-characterized, distal CREs in plants (Supplementary Fig. 1b,c). The number of ACRs within each species ranged from 14,978 to 32,292 (Fig. 1c, Supplementary Tables 1–13) and the number correlated well with the number of annotated genes within each species (Supplementary Fig. 1d–f). ACRs were highly enriched at transcriptional start and end sites, depleted for cytosine DNA methylation[43] and exhibited greater GC content compared with nearby intergenic sequences[44] (Supplementary Fig. 2a–c).

Although genome size varied greatly among these species, the total length of sequence occupied by ACRs was consistent and did not scale linearly with genome size (~5.7–12.8 Mb, accounting for around 0.22–6.5% of the total genome sizes; Fig. 1d,e). Interestingly, there is a significant difference in the location of ACRs relative to genes in plants with different genome sizes, that is, significant factions of ACRs are distributed further away from genes in larger genomes (Fig. 1f,g). We therefore categorized ACRs on the basis of their distance to the nearest gene as genic (gACRs; overlapping a gene), proximal (pACRs; within 2 kb of a gene) or distal (dACRs; >2 kb from a gene). An increase in dACRs accompanied by a loss of pACRs is clearly correlated with increasing genome sizes (Fig. 1h). For example, *A. thaliana* (~119 Mb) and *Spirodela polyrhiza* (~143 Mb) possess the smallest genome sizes in this study, and have 958 and 907 dACRs, which represent approximately 5.9% and 5.7% of all ACRs, respectively. By contrast, *Z. mays* (~2,124 Mb) and *Hordeum vulgare* (~4,834 Mb) have the largest genome sizes in this study, and dACRs represent ~32.8% and ~45.9% of all ACRs in these species, respectively. Taken together, these results show that dACRs are prevalent in plants and are particularly abundant in species with larger genomes. Furthermore, although the total numbers of dACRs do not scale linearly with genome sizes, the proportion of dACRs is correlated with genome size.

[1]Department of Genetics, University of Georgia, Athens, GA, USA. [2]Department of Plant Biology, University of Georgia, Athens, GA, USA.
*e-mail: xiaoyu@uga.edu; schmitz@uga.edu

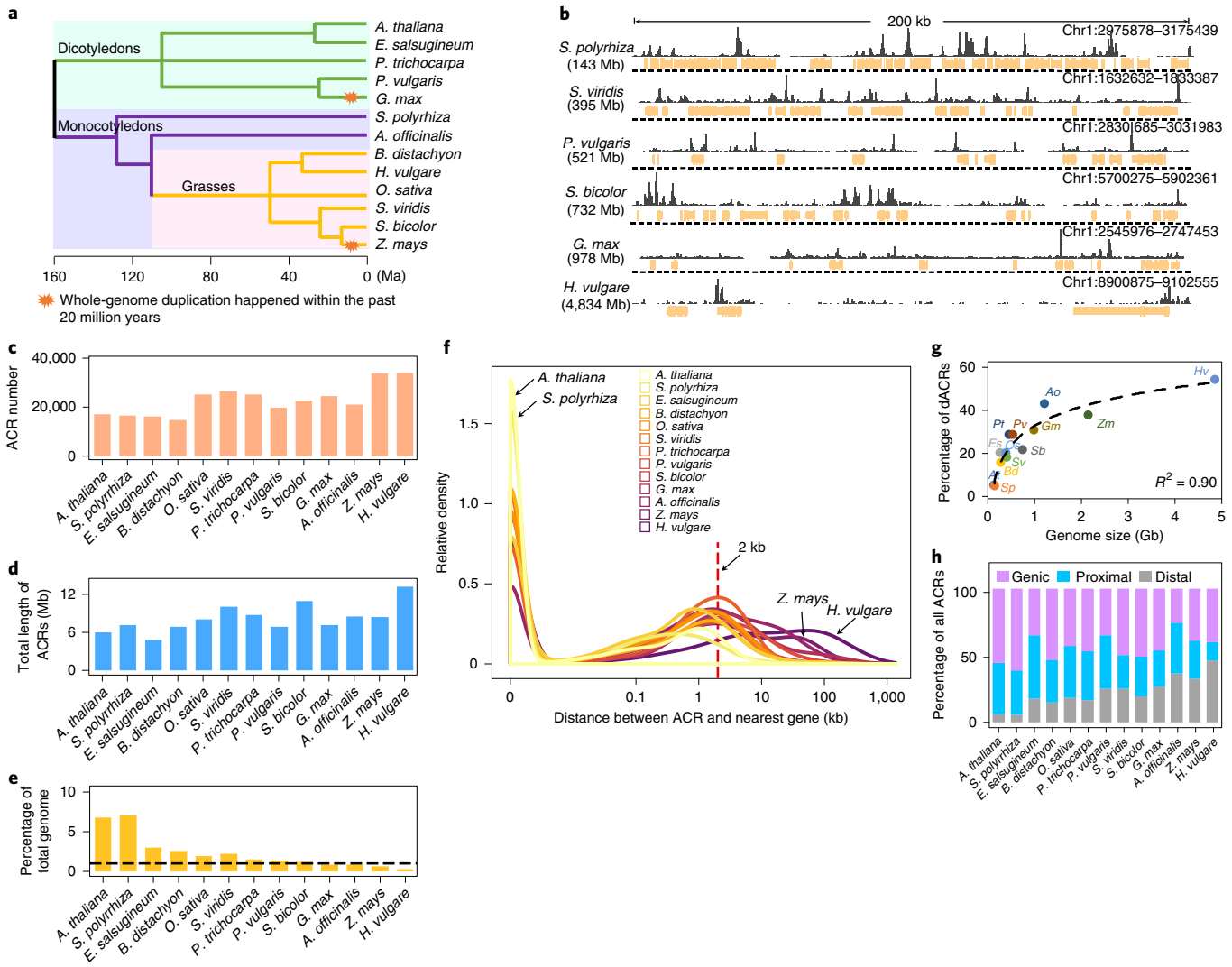


**Fig. 1 | The prevalence of distal accessible regions is a consequence of genome size. a**, Phylogenetic distribution of the plant species that were investigated. **b**, A view of ACRs at a resolution of 300 kb within and between a subset of the plant species studied. The numbers in the parentheses indicate the genome sizes of each species. **c**, The number of ACRs per species (species are ordered on the basis of reference-genome size and not on the basis of phylogenetic distribution). **d**, Total length of sequence possessing an ACR. **e**, The proportion of genome space that possess ACRs. The black dashed line indicates 1%. **f**, A frequency distribution of ACRs and their distance to the nearest genes. **g**, Reference genome size versus the percentage of distal ACRs. The $R^2$ values represent the percentage of the variation that fits a logarithmic model; $n = 13$; $P$ values were calculated by comparing a fitted model to a null model using the 'anova' function; $P < 1 \times 10^{-16}$. *Ao*, *Asparagus officinalis*; *At*, *A. thaliana*; *Bd*, *Brachypodium distachyon*; *Es*, *Eutrema salsugineum*; *Gm*, *G. max*; *Hv*, *H. vulgare*; *Os*, *Oryza sativa*; *Pt*, *Populus trichocarpa*; *Pv*, *P. vulgaris*; *Sb*, *Sorghum bicolor*; *Sp*, *S. polyrhiza*; *Sv*, *Setaria viridis*; *Zm*, *Z. mays*. **h**, The proportion of ACRs that are categorized as genic, proximal or distal within each species investigated.

**Distal accessible chromatin is evolutionarily conserved.** The *Z. mays tb1* and *vgt1* distal control regions are characterized by significantly reduced mutation rates compared with the flanking regions[17,18]. If plant dACRs contain CREs, we expect that dACR sequences should be under purifying selection. Indeed, for 9 of the 13 species, sequence diversity data are available[45,46] and, using these data, we show that all 9 species had considerably lower SNP diversity at dACRs compared with flanking intergenic regions (Fig. 2a). To determine the interspecific conservation of dACR sequences, we identified conserved non-coding sequences (CNSs)[47] within either the dicots species or the monocot grass species in this study (Fig. 2b) and observed a strong enrichment of CNSs at dACRs (Fig. 2c). We further analysed dACRs in two pairs of species that have diverged relatively recently (~12.2 million years ago (Ma) for *Sorghum bicolor* (~732 Mb) and *Z. mays*, and ~24 Ma for *Phaseolus vulgaris* (521 Mb) and *Glycine max* (~965 Mb)). More than half of all dACRs that are present in each of the four species are located in syntenic regions between each species pair (Fig. 2d), making it possible to identify orthologous dACRs by sequence similarity and/or chromatin accessibility. We focused on three scenarios—dACRs that match in sequences and are accessible in both species, dACRs that match in sequence but are only accessible in one species and dACRs that are present in only one species (Fig. 2e). The majority of dACR sequences are present in the sister species, of which approximately two-thirds are also accessible (Fig. 2d). Approximately one-third of sequences that underlie dACRs were not found in their respective sister species (Fig. 2d). By contrast, less than 20% of the inaccessible control regions from the mappable intergenic regions were present in sister species (Fig. 2d). In summary, the vast majority of dACR sequences are under strong

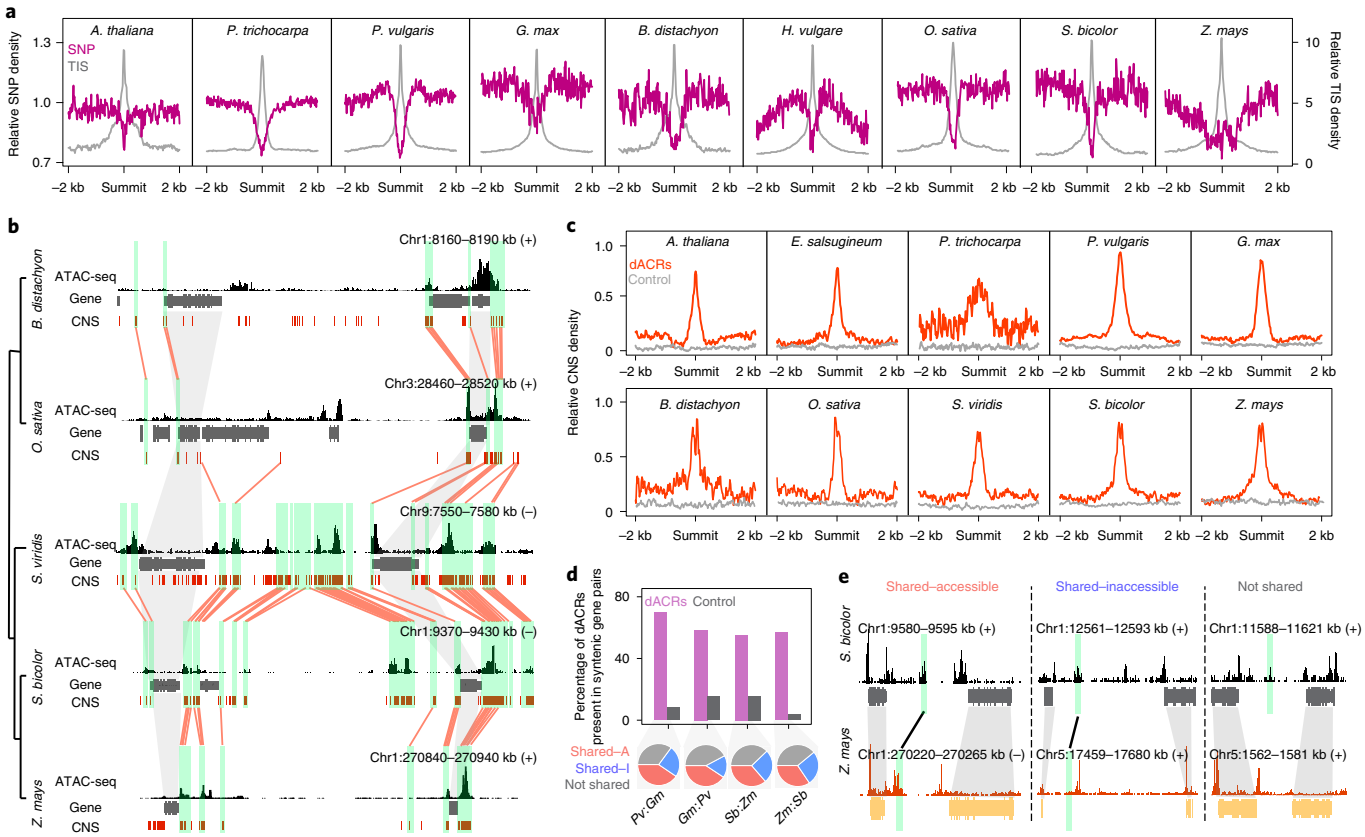


**Fig. 2 | dACRs are preserved between species. a**, The proportion of nucleotide substitutions at dACRs is relatively lower than the surrounding inaccessible regions for the species tested. **b**, An example of a syntenic region containing conserved CNSs and ACRs between the five grass species examined. The vertical lines indicate a CNS. The green transparent boxes show regions that share CNSs. The red lines that bridge green boxes between species indicate CNSs that are found across species. The green shaded regions demarcate an orthologous gene across all five grass species. TIS, Tn5 integration sites. **c**, Distal ACRs are highly enriched in CNSs compared with the intergenic control regions of non-dACRs. **d**, The percentage of sequences that underlie dACRs that are conserved in a syntenic region with the indicate species. The percentage of dACRs with at least one homologous sequence are shown at the top. The percentage of dACR-homologous ACR (hACR) pairs and the proportion that possess shared and/or accessible sequences are displayed in the pie charts for each species pair. Shared-A, shared and accessible; Shared-I, shared and inaccessible. **e**, Representative cases showing variations in dACR conservation between *S. bicolor* and *Z. mays*. The black line links homologous sequences that are present in *Z. mays* (hACRs). Three groups are defined on the basis of the conservation of the sequences and accessibility (shared-accessible, shared-inaccessible and not shared).

purifying selection between species, possibly to retain the sequence specificities for TF binding.

**Transposon content shapes the accessible chromatin landscape.** A major factor that could have contributed to the strong correlation between genome sizes and the distances between ACRs and genes is the differential activities of TEs (Fig. 1), because a higher rate of TE proliferation between ACRs and their target genes could lead to both larger overall genome size as well as the distal localization of ACRs[48]. We tested this possibility using orthologous dACR–gene pairs from *Z. mays* and *S. bicolor*. The overall distance between dACRs and genes was indeed larger in *Z. mays* and, importantly, the difference in distances could be largely explained by the amount of TE-derived sequences (Fig. 3a–d). A similar correlation was observed between *G. max* and *P. vulgaris*, suggesting that TE expansion is a general feature that contributes to the presence of dACRs (Fig. 3e,f). We further found that the variation in distance between dACRs and genes did not affect the expression of nearby genes (Supplementary Fig. 3a). In addition to moving ACRs away from genes, TE activities could also create new dACRs in individual species when regulatory elements contained by certain TEs are introduced into intergenic regions by species-specific transposition events[18,49–52]. Indeed, although dACRs are generally depleted from TEs (Supplementary Fig. 3b), species-specific dACRs are strongly enriched within TEs (Fig. 3g,h, Supplementary Fig. 3c), particularly in class II TEs (permutation test, $P < 1 \times 10^{-4}$; Fig. 3i), a feature shared by all of the plant species studied here.

A second major source of variation in the size of plant genomes results from polyploid events that are followed by differential fractionation[53,54]. Although these events are not expected to directly alter the distance between ACRs and their target genes, the paralogous ACR–gene pairs produced by past events provide an excellent opportunity to understand the emergence, retention and movement of dACRs through evolutionary time. We studied this question using dACRs identified in *G. max* and *Z. mays* as both species experienced polyploidization followed by differential fractionation (Fig. 1a). Although these events in *G. max* (~12.4 Ma)[55] and *Z. mays* (~5–12 Ma)[54] occurred on similar timescales, the majority of dACRs are retained in *G. max* in comparison to *Z. mays* (Fig. 3j,k). The *Z. mays* genome has experienced greater degrees of fractionation compared with *G. max*, consistent with the rates of retention of genes within each of these species following past polyploidy events. Of the sequences that underlie dACRs that are shared within *G. max*, the majority of them are accessible; by contrast, in *Z. mays*, roughly half are still accessible (Fig. 3k). Similar to the species-specific dACRs, sub-genome-specific dACRs are also strongly

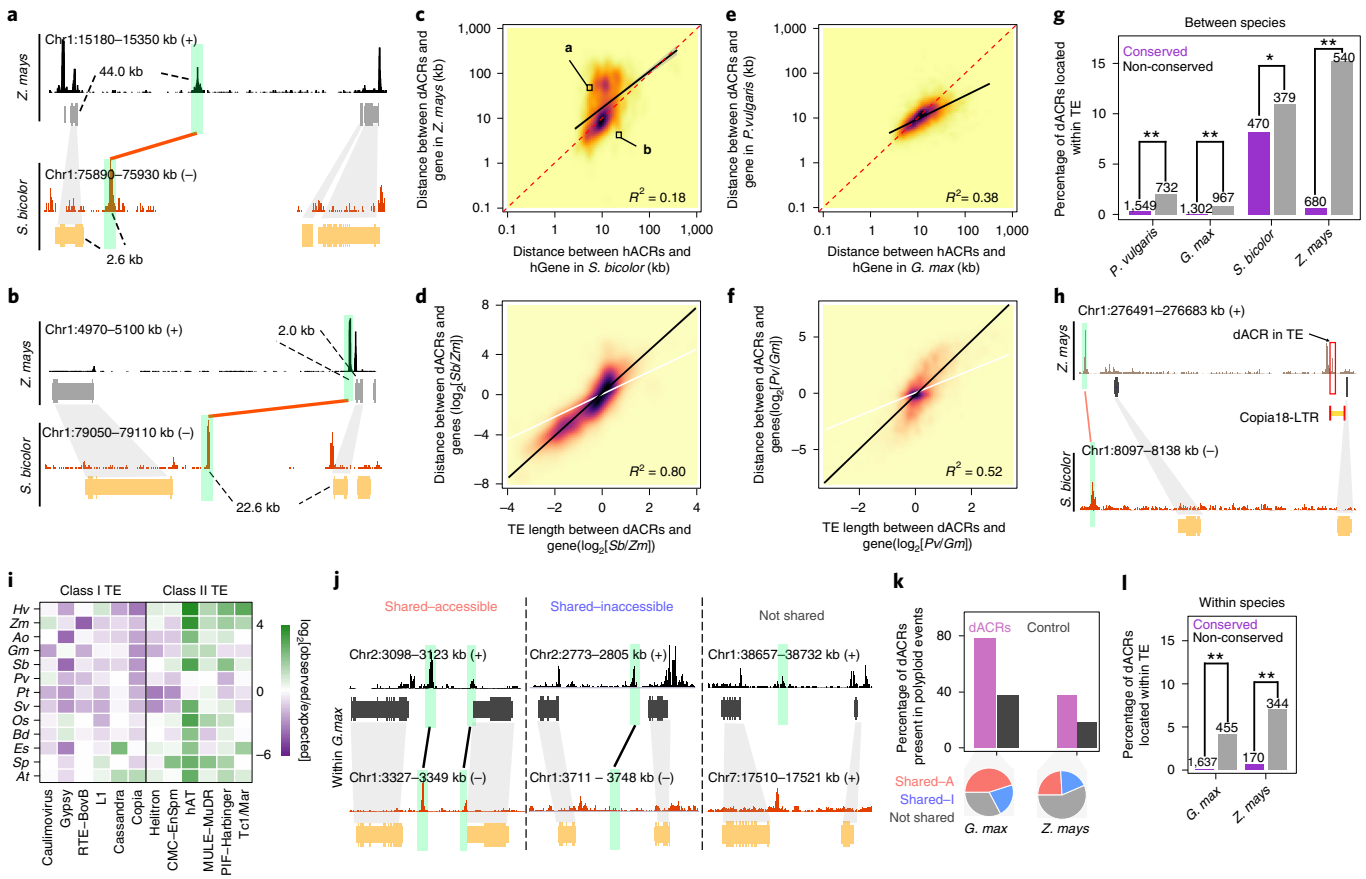


**Fig. 3 | TEs play important roles in the distribution of ACRs across plant genomes. a**,**b**, Two examples of independent genomic locations of variation in dACR-to-gene distances between *Z. mays* and *S. bicolor*. The green transparent boxes linked by red lines demarcate ACRs with high sequence homology. **c**, Statistical results of the dACR-to-gene distance between *Z. mays* and *S. bicolor.* The characters 'a' and 'b' indicate the examples shown in **a** and **b**, respectively. The $R^2$ values represent the percentage of the variation that fits a linear model; $n = 358$; *P* values were calculated by comparing a fitted model to a null model using the 'anova' function; $P < 1 \times 10^{-16}$. The red dotted line indicates that the distances are equal. **d**, A comparison of dACR-to-gene distances with the number of annotated TE sequences present in *Z. mays* and *S. bicolor.* The $R^2$ values represent the percentage of the variation that fits a linear model; $n = 358$; *P* values were calculated by comparing a fitted model to a null model using the 'anova' function; $P < 1 \times 10^{-16}$. The white line indicates that the ratios are equal. The shaded regions indicate the standard normal density. **e**,**f**, Comparison of dACR-to-gene distances between *G. max* and *P. vulgaris* (**e**) and the number of annotated TE sequences present in each genome (**f**). The $R^2$ values represent the percentage of the variation that fits a linear model; $n = 948$; *P* values were calculated by comparing a fitted model to a null model using the 'anova' function; $P < 1 \times 10^{-16}$. The red dashed line indicates that the distances are equal and the white line indicates that the ratios are equal. **g**, The proportion of overlap between TEs and different conservation classes of dACRs. 'Conserved' indicates dACRs with at least one hACR(s) between species; 'non-conserved' indicates species-specific dACRs. The numbers indicate the sample sizes used in the analysis. $^{**}P < 0.01$, $^{*}P < 0.05$; the statistical analysis was performed using two-sided Fisher's exact tests. **h**, A representative region highlighting the co-localization of a dACR and annotated long terminal repeat (LTR) that is present in *Z. mays* (red outlined box) but absent in *S. bicolor*. The green transparent boxes linked by red lines demarcate ACRs with high sequence homology. **i**, The $\log_2$-transformed ratio of observed versus expected proportions of dACR-containing TEs across different TE families for each of the 13 angiosperm genomes. **j**, Representative cases showing variation in dACR conservation within *G. max* following whole-genome duplication. The black line links homologous sequences that are present in duplicated events (hACRs). Three groups are defined on the basis of the conservation of the sequences and accessibility (shared-accessible, shared-inaccessible and not shared). **k**, Conservation of sequence and chromatin accessibility for dACRs from duplicated regions of *G. max* and *Z. mays*. Top, the proportion of dACRs with at least one intra-genomic homologous sequence. Bottom, the relative distribution of shared-accessible, shared-inaccessible and not shared dACR–hACR pairs for *G. max* and *Z. mays*. **l**, A comparison of TE-overlapping rates for conserved and non-conserved dACRs within species. The numbers indicate the sample sizes used in the analysis. $^{**}P < 0.01$; the statistical analysis was performed using two-sided Fisher's exact tests.

enriched within TEs (Fig. 3l). The polyploid genome in *Z. mays* is ancient, and biased gene loss and expression continues at present[54]. Consistent with previous reports studying epigenomic signatures in *Z. mays*, we found that most of the subgenome-specific dACRs were located within the A genome (Supplementary Fig. 3d). Consistent with greater TE activity in *Z. mays*, we found that the distance between dACRs and genes among sub-genomes was more diverse in *Z. mays* than *G. max* (Supplementary Fig. 4a,b). As a consequence, variation in the distance of dACRs to genes between whole genome duplication (WGD) segments was largely accounted for by TE sequences (Supplementary Fig. 4c,d) and the distance of ACRs to the nearby gene did not affect their expression differentially compared with those ACRs that were close (Supplementary Fig. 4e,f). Collectively, these results show that WGD and TE activity shape the landscape of ACRs in a species-specific manner.

**Plant dACRs are characterized by three distinct chromatin states.** In mammalian genomes, distal CREs are demarcated by H3K4

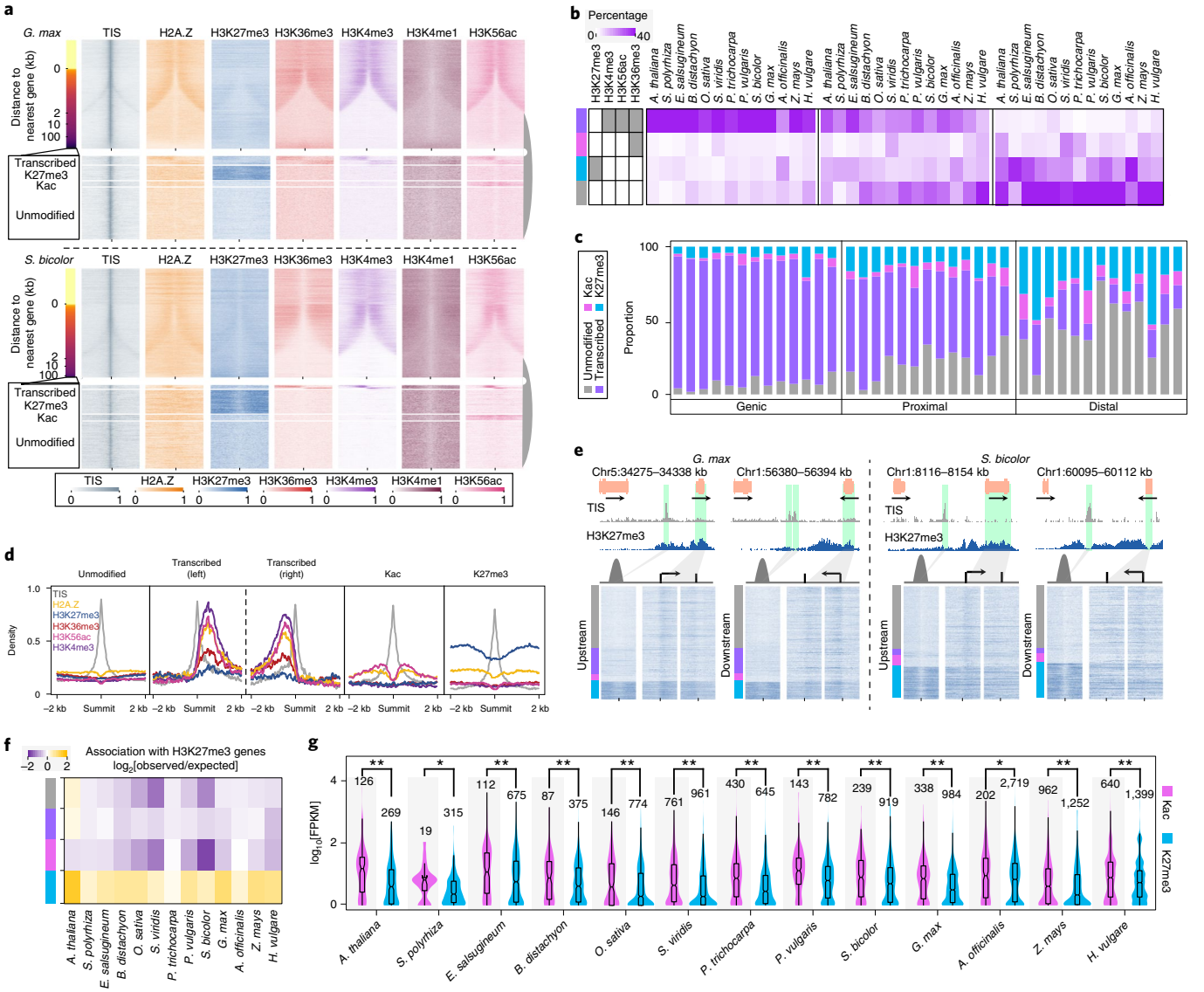


**Fig. 4 | ACRs are characterized by distinct and conserved chromatin states depending on their distance to genes. a**, The chromatin state of ACRs is dependent on distance to genes. ACRs are arranged by their distance to the nearest gene. The results for *G. max* and *S. bicolor* are shown as representative examples. The histone modifications of dACRs are shown in the bottom panel for each species. **b**, The percentage of ACRs flanked by four major histone-modification combinations in flowering plant genomes are distinguished by their distance from gene bodies. The colour key in **c** applies to **b**, **e** and **f**. **c**, The relative proportions of transcribed, unmodified, Kac and K27me3 classifications for ACRs coinciding various genomic features. 'Transcribed' indicates ACRs with H3K4me3 and/or H3K36me3 modifications; 'K27me3' indicates ACRs with only H3K27me3 modifications; 'Kac' indicates ACRs with only H3K56ac modifications; 'Unmodified' indicates ACRs without the any of these four histone modifications. **d**, Metaplots of histone modification and H2A.Z profiles at dACRs. The left and right transcribed groups represent unannotated transcribed regions ordered on the basis of the direction of transcription. **e**, Comparison of chromatin states for dACRs and nearby genes. In this example, H3K27me3 data for dACRs located upstream and downstream of nearby genes are shown for *G. max* (left) and *S. color* (right). **f**, The proportion of dACR classifications that are enriched or depleted near H3K27me3 genes relative to control regions from 100 permutations. **g**, Expression levels of nearby genes flanked by H3K27me3 or H3K56ac dACRs. The box plot displays the distribution of data on the basis of the five number summary: minimum, first quartile, median, third quartile and maximum. The central rectangle spans the first quartile to the third quartile (the interquartile range), the segment inside the rectangle shows the median and the whiskers indicate the minimum and maximum values. The brown shadows indicate the distribution. The numbers indicate the sample size used in the analysis. $^{**}P < 0.01$; the statistical analysis was performed using two-sided Wilcoxon ranked sum tests.

monomethylation (me1)[56,57]. Moreover, the depletion of cytosine DNA methylation in conjunction with the enrichment of H3K27 acetylation (ac) is predictive of active enhancers[43,58]. However, H3K4me1 is not associated with distal enhancers in *A. thaliana* or *Z. mays*, whereas H3K9ac and H3K27ac are associated with putative enhancers in plants[22,24,28,32,33]. To determine whether plant dACRs were associated with specific chromatin features in the 13 species, we produced genome-wide maps of histone modifications (H3K4me1, H3K4me3, H3K36me3, H3K56ac, H3K27me3 and H3) and the histone variant H2A.Z in all 13 species. Genic chromatin modification features are similar between all 13 species and to that of previously published plant epigenomes (Supplementary Fig. 5)—expressed genes are enriched in H3K4me3, H3K56ac and H2A.Z at the transcriptional start site, and enriched in H3K4me1 and H3K36me3 in

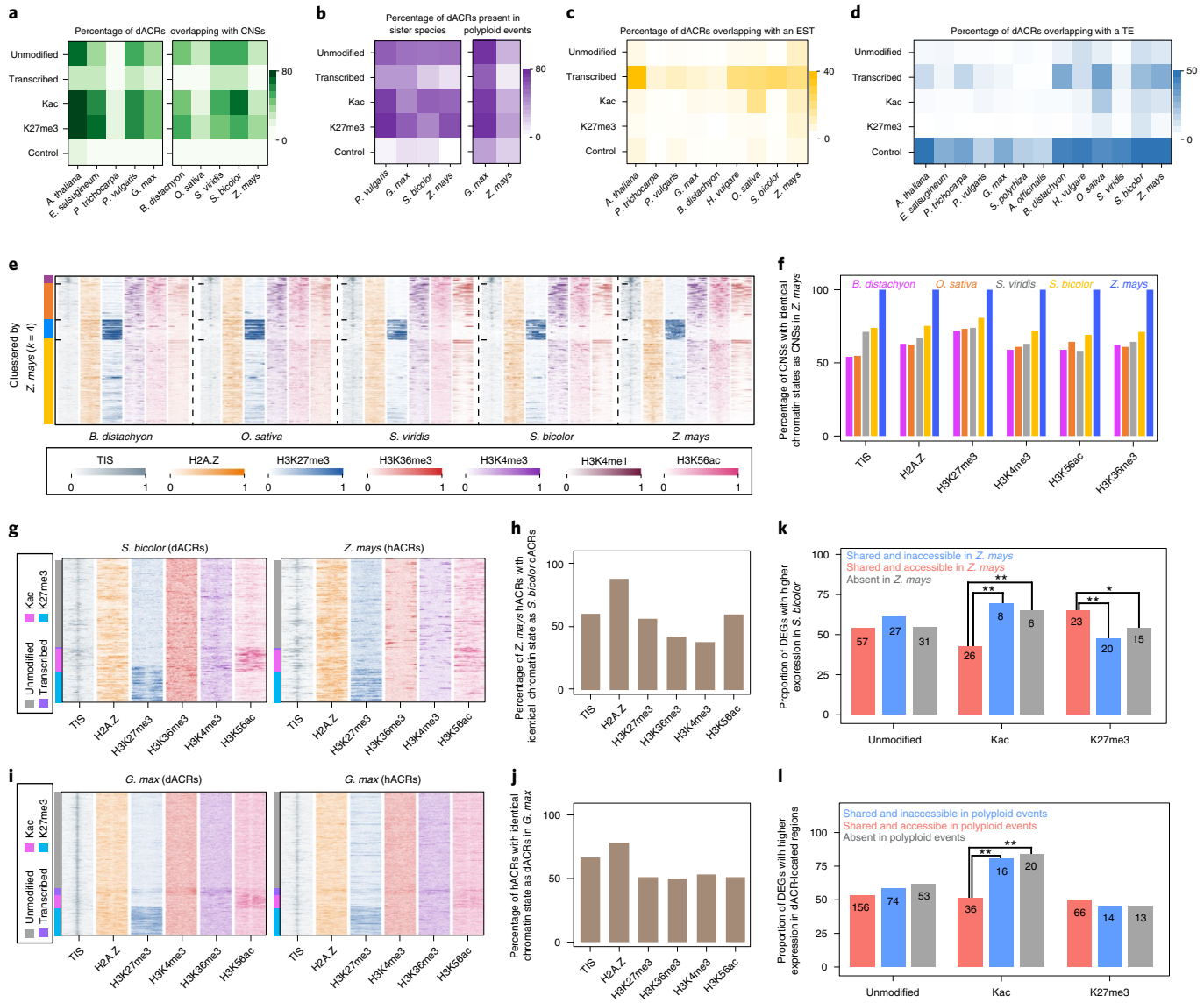


**Fig. 5 | The chromatin states of dACRs are conserved between species. a**, The proportion of dACRs that coincide with CNSs across dACR groups and species. **b**, The proportion of dACRs that are conserved with a sister species (left) and between whole-genome-duplicated sequences within species (right). **c**, The overlapping proportions between different dACR groups and expression sequence tags (ESTs) among species. **d**, The proportion of dACRs contained within TE sequences for the various dACR classifications in each species. **e**, A comparison of chromatin states in 4 kb windows that are centred on CNSs shared among all grass species. A *k*-means test ($k = 4$) in *Z. mays* was performed to cluster and order CNSs in the other species. Each coloured box indicates a distinct *k*-means group. **f**, The percentages of CNSs with matching chromatin accessibility and histone modifications among monocots relative to *Z. mays*. **g**–**j**, Chromatin accessibility and adjacent chromatin modifications for dACRs and homologous dACRs between related species (**g**) and quantification (**h**), and the same comparison for *G. max* sub-genomes (**i**) and quantification (**j**). Each coloured box in **g** and **i** represents a distcint histone modification group, as indicated in the legend. **k**, Expression levels of *S. bicolor* genes closest to *S. bicolor*-specific dACRs (absent or inaccessible dACR in the corresponding syntenic region of *Z. mays*; Fig. 2e) showed chromatin-state-dependent variations. The numbers indicate the sample sizes used in the analysis. The statistical analysis was performed using two-sided Fisher's exact tests. **l**, Expression levels for genes closest to subgenome-specific dACRs within *G. max* (absent or inaccessible dACR in the corresponding duplicated region in the opposing subgenome of *G. max*; Fig. 4j) showed chromatin-state-dependent variations. The numbers indicate the sample sizes used in the analysis. $^{**}P < 0.01$; the statistical analysis was performed using two-sided Fisher's exact tests.

gene bodies (Supplementary Fig. 5). By contrast, repressed genes are enriched in H3K27me3 and H2A.Z (Supplementary Fig. 5). ACRs were next categorized into classes on the basis of their local chromatin modifications. ACRs characterized by H3K4me3, H3K56ac and H3K36me3 (transcribed) are indicative of active transcription and are enriched in genic and proximal ACRs (Fig. 4a–c, Supplementary Fig. 6). However, with increasing distance from the nearest annotated gene, the chromatin modification at proximal and distal ACRs in all species changes to an unmodified state (unmodified), a H3K56ac state (Kac) or a H3K37me3 state (K27me3), or remains in a transcribed chromatin state (Fig. 4a–c, Supplementary Fig. 6). Importantly, in contrast to in mammalian genomes, H3K4me1 is not commonly associated with dACRs in any of the 13 species (Fig. 4a, Supplementary Fig. 6). Most of the dACRs are present in an unmodified state, although the histone variant H2A.Z is enriched at these regions (Fig. 4d). The Kac dACRs are usually associated

with higher expression levels of nearby genes and may function as enhancers. By contrast, the K27me3 dACRs are probably associated with the Polycomb silencing pathways and represent distal repressor elements[59,60] (Fig. 4d). Genes flanking K27me3 dACRs are more likely to be also associated with H3K27me3 and lower levels of expression (Fig. 4e–g). Importantly, Gene Ontology classification revealed that genes flanking the unmodified, Kac and K27me3 dACRs are enriched in terms related to transcriptional regulation, whereas those flanking the transcribed group lacked significant functional enrichment (Supplementary Fig. 7). Chromatin can form long-range loops that bring together dACRs with cognate genes. Indeed, the expression status of target genes predicts the chromatin state at the physically coupled dACRs—highly expressed genes interact with Kac dACRs and genes with low expression interact with K27me3 dACRs[61]. These interactions, which are empirically demonstrated to occur in *Z. mays*, probably occur within the diverse angiosperms that we profiled in this study. These data support that dACRs possess distinct chromatin environments and that, with the exception of the transcribed group, they are similar to their probable target genes and likely key determinants that modulate transcription.

**Chromatin states of dACRs are conserved between species.** The conservation of dACR function between species varies depending on the local chromatin environment. For example, the transcribed group of dACRs is a relatively large class that is associated with H3K4me3, H3K56ac and H3K36me3, resembling expressed genes (Fig. 4d). Further inspection of these dACRs revealed that they were most variable in sequence among species (Fig. 5a,b) and most were putative promoters for unannotated genes or long noncoding RNAs (Fig. 5c, Supplementary Fig. 8). Interestingly, 4–34% of transcribed dACRs were located within TE sequences (Fig. 5d), possibly underlying sequences that are capable of activating TE transcription. Importantly, histone modifications at orthologous dACRs are conserved among species. For example, when *Z. mays* dACRs were clustered on the basis of histone modifications, the orthologous dACRs in other species showed strikingly similar profiles (Fig. 5e–h). Similar results were found within comparisons among the eudicot species that we studied (Supplementary Fig. 9) and among the paralogous dACRs produced by WGDs (Fig. 5i,j). However, we observed changes in gene expression for dACRs exhibiting species specificity. For example, K27me3 dACRs in *S. bicolor* that were inaccessible or absent in *Z. mays* were concomitant with lower expression of nearby genes (Fig. 5k). Reciprocally, Kac dACRs present in only one species were associated with higher expression (Fig. 5k). Similar patterns of activation and repression were found in paralogue-specific dACRs within a species (Fig. 5l). Moreover, we found that genes associated with species- or subgenome-specific dACRs had considerably more variable expression relative to genes with shared dACRs (Supplementary Fig. 10). Taken together, these results show that plant dACRs are associated with three major types of chromatin modification features that are distinct from mammalian distal CREs and are probably functionally important.

## Discussion

The results from decades of genetic and transformation studies in *A. thaliana* were consistent with recent genome-wide chromatin accessibility data in that the vast majority of CREs in its small genome were located adjacent to the transcription start sites[2]. Although several individual examples of long-distance transcriptional regulation have been described in other plants, the prevalence[15–22], evolutionary dynamics and chromatin features of plant distal CREs remain unclear. Through genome-wide comparisons of DNA sequences, chromatin accessibility and histone modifications of 13 plant species, we show that distal ACRs are prevalent in plants and are abundant in larger and more complex genomes. On the basis of functional studies in maize of these distal ACRs, it is predicted that many of them possess bona fide CREs indicating that many ACRs are CREs[61]. These distal CREs are highly conserved within species and, in many cases, among species. The presence of distal CREs and their distances from target genes are dynamic and associated with TEs activities. Interestingly, chromatin features that are commonly found at metazoan distal CREs seem to be missing in plants; instead, plant distal CREs are either unmodified, or contain histone acetylation or H3K27me3. Furthermore, the data presented here provide a rich resource for the understanding of the regulation of individual genes as well as regulatory changes that underlie phenotypic evolution in plants. Finally, the identification of distinct chromatin features at plant distal CREs provides the starting point to understand how TFs interact with chromatin pathways to regulate plant gene expression and development.

## Methods

**Plant material and growth conditions.** *A. thaliana* (accession Col-0) was grown on 1/2 MS plates at 25 °C under continuous light for 7 d. Leaves were collected for the experiments. *S. polyrhiza* (accession 7498) were grown in liquid medium (Schenk and Hildebrandt basal salt mixture) at 25 °C under continuous light. Whole seedlings were collected 3 d after transferring to a new medium. *E. salsugineum* (accession Shandong) and *A. officinalis* (accession Gijnlim) were grown in the soil for approximately 10 d under 25 °C and photoperiodic lighting (16 h light–8 h dark). All of the young leaves were collected for the experiments. *P. vulgaris* (accession G19833) and *G. max* (accession William82) were grown in the soil for approximately 10 d at 25 °C under photoperiodic lighting (16 h light–8 h dark). The second and third leaves were collected for the experiments. *B. distachyon* (accession Bd21), *O. sativa* (accession Nipponbare), *S. viridis* (accession A10), *S. bicolor* (accession BTx623), *Z. mays* (accession B73) and *H. vulgare* (accession Morex) were grown in soil for around 6–7 d at 25 °C under 16 h light–8 h dark. The inner second leaves—which contains the third and fourth leaves sheathed inside—were used for experiments. *P. trichocarpa* (accession Stettler) leaves were collected from the tree and flash-frozen with liquid $N_2$ immediately for the subsequent experiments.

**ATAC-seq.** ATAC-seq was performed as described previously[41]. For each replicate, approximately 200 mg of freshly collected leaves or flash-frozen leaves were immediately chopped with a razor blade in approximately 1 ml of prechilled lysis buffer (15 mM Tris-HCl pH 7.5, 20 mM NaCl, 80 mM KCl, 0.5 mM spermine, 5 mM 2-mercaptoethanol and 0.2% Triton X-100). The chopped slurry was filtered twice through Miracloth and once through a 40 µm filter. The crude nuclei were stained with 4,6-diamidino-2-phenylindole (DAPI) and loaded into a flow cytometer (Beckman Coulter, MoFlo XDP). Nuclei were purified by flow sorting and washed as described previously[41]. The sorted nuclei were incubated with 2 µl Tn5 transposomes in 40 µl tagmentation buffer (10 mM TAPS-NaOH pH 8.0, 5 mM $MgCl_2$) at 37 °C for 30 min without rotation. The integration products were purified using a Qiagen MinElute PCR Purification Kit or NEB Monarch DNA Cleanup Kit and then amplified using Phusion DNA polymerase for 10–13 cycles. PCR cycles were determined as described previously[62]. Amplified libraries were purified using AMPure beads to remove primers.

**RNA-seq.** The collected leaves were flash-frozen with liquid $N_2$ immediately after collection. The samples were ground to a powder using a pestle and mortar in liquid $N_2$. Total RNA was extracted and purified using TRIzol Reagent (Thermo Fisher Scientific) following the manufacturer's instructions. For each sample, 1.3 µg of total RNA was prepared for sequencing using the Illumina TruSeq mRNA Stranded Library Kit (Illumina) following the manufacturer's instructions.

**Chip-seq.** Chromatin immunoprecipitation (ChIP) was performed using a previously described protocol[59]. In brief, approximately 1 g freshly harvested leaves or flash-frozen leaves was chopped into 0.5 mm cross-sections and crosslinked as described previously[59]. The samples were immediately flash-frozen in liquid $N_2$ after crosslinking. Nuclei were extracted and lysed in 300 µl of lysis buffer. Lysed-nuclei suspension was sonicated using a Diagenode Bioruptor on the high setting, 30 cycles of 30 s on and 30 s off. To make antibody-coated beads, 25 µl Dynabeads Protein A (Thermo Fisher Scientific, 10002D) was washed with ChIP dilution buffer and then incubated with around 1.5 µg antibodies (anti-H3K4me1, Abcam, ab8895; anti-H3K4me3, Millipore-Sigma, 07–473; anti-H3, Abcam, ab1791; anti-H3K36me3, Abcam, ab9050; anti-H3K27me3, Millipore-Sigma, 07–449; anti-H3K56ac, Millipore-Sigma, 07–667; anti-H2A.Z, from R. Deal's laboratory) in 100 µl ChIP dilution buffer for at least 1 h at 4 °C. The sonicated chromatin samples were centrifuged at 12,000 g for 5 min and the supernatants were diluted tenfold in ChIP dilution buffer to reduce the SDS concentration to 0.1%. ChIP input aliquots were collected from the supernatants. For all of the samples and replicates, 300–500 µl of diluted chromatin was incubated with the antibody-coated beads at 4 °C for at least 4 h or overnight, then washed, reverse-crosslinked and treated with

proteinase K in accordance with the protocol[59]. DNA was purified using a standard phenol–chloroform extraction method, followed by ethanol precipitation.

The DNA samples were end-repaired using the End-It DNA End-Repair Kit (Epicentre) following the manufacturer's protocol. DNA was cleaned up on AMPure beads (Beckman Coulter) with size selection of 100 bp and larger. The samples were eluted into 43 µl Tris-HCl and subsequently underwent a 50 µl A-tailing reaction in NEBNext dA-tailing buffer with Klenow fragment (3′ -> 5′ exo⁻) at 37 °C for 30 min. A-tailed fragments were ligated to Illumina TruSeq adapters and purified with AMPure beads. The fragments were amplified using Phusion polymerase in a 50 µl reaction following the manufacturer's instructions. The following PCR program was used: 95 °C for 2 min; 98 °C for 30 s; then 15 cycles of 98 °C for 15 s, 60 °C for 30 s and 72 °C for 4 min; and once at 72 °C for 10 min. The PCR products were purified using AMPure beads to remove primers.

**Sequencing information.** Sequencing of ATAC-seq, RNA-seq and ChIP-seq libraries was performed at the University of Georgia Genomics & Bioinformatics Core using an Illumina NextSeq 500 instrument. ATAC-seq samples were sequenced in paired-end 35 bp reads. RNA-seq libraries for *P. trichocarpa* were sequenced with paired-end 75 bp reads and all of the others were sequenced with single-end 75 bp reads. ChIP-seq libraries were sequenced in single-end 75 bp reads. Information on read counts and alignment statistics is provided in Supplementary Tables 14–16.

**ATAC-seq raw data processing and alignment.** Raw reads were trimmed using Trimmomatic v.0.36 (ref. [63]). Reads were trimmed for NexteraPE using a maximum of two seed mismatches, a palindrome clip threshold of 30 and a simple clip threshold of 10. Reads that were shorter than 30 bp were discarded. Trimmed reads were aligned to the reference genome using Bowtie v.1.1.1 (ref. [64]) using the following parameters: 'bowtie -X 1000 -m 1 -v 2 --best –strata'. *A. thaliana*, *E. salsugineum*, *P. vulgaris*, *G. max*, *B. distachyon*, *O. sativa*, *S. viridis*, *P. trichocarpa* and *S. bicolor* genomes were obtained from JGI (phytozome v.11)[45,46]; *Z. mays* from MaizeGDB[65,66]; *H. vulgare L.* from Ensemble (v.42)[67]; *A. officinalis* from the Asparagus genome project (http://asparagus.uga.edu/tripal/)[68]; and *S. polyrhiza* from Michael et al.[69]. Aligned reads were sorted using SAMtools v.1.3.1 (ref. [70]) and clonal duplicates were removed using Picard v.2.16.0 (http://broadinstitute.github.io/picard/).

**RNA-seq raw data processing, alignment and quantification of expression.** Raw reads were trimmed with Trimmomatic v.0.36 (ref. [63]). Reads were trimmed for Trueseq3 barcodes using a maximum of two seed mismatches, a palindrome clip threshold of 30 and a simple clip threshold of 10. Reads shorter than 50 bp were discarded. The remaining reads were aligned to the same version reference genome with ATAC-seq using TopHat v.2.1.1 (ref. [71]) using the following parameters: '--library-type fr-firststrand --max-intron-length 10000'. Gene expression values were computed using Cufflinks v.2.2.1 (ref. [72]). The average fragments per kb of transcript per million mapped reads (FPKM) value of the two replicates was used for the analysis. To change the FPKM to a $\log_2$ scale, 1 pseudo count was added. To compare the expression from different species, transcripts per kb of transcript per million mapped reads (TPM) values were calculated using TPMCalculator[73] and normalized to the average TPM of the gene. Genes with at least twofold changes in expression were used as differentially expressed genes in the subsequent analysis. To avoid overrepresenting the expression changes of low-expression genes, 1 pseudo count was added before calculating the fold change.

**ChIP-seq raw data processing and alignment.** Raw reads were trimmed using Trimmomatic v.0.36 (ref. [63]). Reads were trimmed for Trueseq3 barcodes with a maximum of two seed mismatches, a palindrome clip threshold of 30 and a simple clip threshold of 10. Reads shorter than 50 bp were discarded. For all of the species except *H. vulgare*, the remaining reads were aligned to the same reference genome with ATAC-seq using Bowtie v.1.1.1 (ref. [64]) with the following parameters: '-m 1 -v 2 --best --strata –chunkmbs 1024-S'. For *H. vulgare*, the remaining reads were aligned to the reference genome using Bowtie2 v.2.1.1 (ref. [74]). Reads with MAPQ > 5 were used for the subsequent analysis. Aligned reads were sorted using SAMtools v.1.3 and duplicated reads were removed using Picard v.2.16.0 (http://broadinstitute.github.io/picard/).

**Identification of ACRs.** MACS2 (ref. [75]) was used to define ACRs with the '--keep-dup all' function. To find high-quality ACRs, the following filtering steps were generally performed: (1) ACRs called with MACS2 were split into 50 bp windows with 25 bp steps; (2) the Tn5 integration frequency in each window was calculated and normalized to the average frequency in the total genome; (3) windows passing the integration frequency cut-off were merged together with 150 bp gaps; (4) small regions with only one window were then filtered with 'length > 50 bp'; (5) regions aligning to the mitochondrial or chloroplast genome from NCBI Organelle Genome Resources were also removed. The sites within ACRs with the highest Tn5 integration frequency were defined as summits.

**Definition of intergenic negative control regions.** To create the intergenic negative control regions, we first identified the uniquely mappable regions by remapping all of the possible simulated 75 bp fragments from each reference genome, using the same parameters as for ChIP-seq analysis. Genomic regions with mapped reads were considered to be uniquely mappable. Annotated genes and their 2 kb nearby regions, as well as gene-distal ACRs, were removed. Negative control regions with the same length distribution to dACRs were then generated using the 'shuffle' command in BEDTools[76].

**Identification of conserved CNSs.** The CNS Discovery Pipeline v.3.0 (ref. [47]) was used to find the CNSs by pairwise comparison. The genomic sequences and annotations (BED format), which were the same version for the reference genome for ATAC-seq, RNA-seq and ChIP-seq, were used as input. Every two genomes within the dicots or Poaceae (excluding *H. vulgare*) category were selected for the analysis. Common CNSs, which were 2 kb away from genes in at least one genome, were used to determine chromatin accessibility and histone modifications.

**Identification of syntenic regions and homologous ACR sequences.** The inter-species syntenic genes and intraspecies duplicated genes were identified using the Multiple Collinearity Scan toolkit (MCscanX)[77] with the following parameters: '-m 10 -w 4'. The longest annotated protein peptides and gene coordinates of each gene were used as input. The output co-linearized genes were considered to be syntenic (or duplicated) genes and were used in the subsequent analysis. To determine whether an ACR was present in syntenic (or duplicated) regions, the two genes directly nearby, for which syntenic (or duplicated) genes were also next to each other in sister species or other sub-genomes, were used for the analysis. ACR sequences were then BLAST-aligned to regions including (1) the intergenic sequences between the syntenic (or duplicated) gene pair; (2) two syntenic gene sequences, by NCBI+ BLAST[78] with the following parameters: '-task blastn-short -evalue 1e-3 -max_target_seqs 10'.

**Heat map and metaplot analysis.** To generate signal densities for ATAC-seq and ChIP-seq, each sample was normalized to their input samples by adding 1 pseudo count to each 10 bp window, respectively. The signal densities for ATAC-seq, DNA methylation[79], SNPs[45,46], CNSs and histone ChIP-seq from 2 kb upstream to 2 kb downstream around ACR summits, annotated transcription start sites, annotated transcription end sites, CNS centre or homologous region centre were then calculated using deepTools v.3.0.2 (ref. [80]) with the following parameters: '-a 2000 -b 2000 -bs 20'. The matrix was then normalized to 0–1 with the 98th quantile value of each sample as an upper limit. The average values of each bin were used to construct metaplots.

**Identification of ACR clusters by *k*-means clustering.** Signal densities for the histone modifications H3K27me3, H3K36me3, H3K4me3 and H3K56ac ChIP-seq from 2 kb upstream to 2 kb downstream around ACR summits were calculated using deepTools v.3.0.2 (ref. [80]) with the following parameters: '-a 2000 -b 2000 -bs 20'. The matrix was then normalized to 0–1 with 98th quantile value of each sample as an upper limit. Then, *k*-mean tests were used to determine whether the ACRs were linked to the histone modification. The number of clusters were determined by Elbow, Silhouhette and Gap statistic methods[81], and subsequent manual inspection. ACRs with every combination of the four histone modifications were grouped into four main groups.

**Annotation of TEs.** Repetitive elements in each species were first identified using RepeatMasker v.4.07 (ref. [82]) with the following parameters: '-species "angiosperms" -pa 10 -no_is -gff -e wublast'. TEs frequently insert into one another, making the identification of complete elements challenging. To overcome this, the genomic coordinates of transposon fragments output from RepeatMasker that were derived from the same alignment were merged.

**Analysis of dACR enrichment within TE superfamilies.** The relative frequency of dACR enrichment within various TE superfamilies was determined by estimating the proportion of TEs that harbour a dACR. dACR coordinates were required to be completely encompassed by the underlying TE interval. To assess dACR enrichment, simulated random mappable regions (using the 75 nucleotide mappable regions as described above for BEDtools shuffle) composed of the same interval lengths and number of sites were constructed and compared with each TE superfamily similar to dACRs. The null distribution of TE overlap for each TE superfamily was determined by 10,000 permutations. Empirical *P* values were determined by comparing the observed proportion of TEs containing ACRs with the null distribution. The heat map depicting observed versus expected $\log_2$-transformed fold changes was constructed using the TE-ACR proportions for observed over the permutation-derived mean.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

The data generated in this study have been uploaded to the Gene Expression Omnibus (GEO) database and can be retrieved through accession number GSE128434. The data from this study can also be viewed interactively on the publicly accessible epigenome browser at http://epigenome.genetics.uga.edu/PlantEpigenome/.

Received: 8 May 2019; Accepted: 9 October 2019;
Published online: 18 November 2019

## References


1. Narlikar, G. J., Fan, H. Y. & Kingston, R. E. Cooperation between complexes that regulate chromatin structure and transcription. *Cell* **108**, 475–487 (2002).
2. Priest, H. D., Filichkin, S. A. & Mockler, T. C. *cis*-Regulatory elements in plant cell signaling. *Curr. Opin. Plant Biol.* **12**, 643–649 (2009).
3. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.* **20**, 207–220 (2019).
4. Sakabe, N. J., Savic, D. & Nobrega, M. A. Transcriptional enhancers in development and disease. *Genome Biol.* **13**, 238 (2012).
5. Thurman, R. E. et al. The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
6. Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
7. Mishra, A. & Hawkins, R. D. Three-dimensional genome architecture and emerging technologies: looping in disease. *Genome Med.* **9**, 87 (2017).
8. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* **15**, 272–286 (2014).
9. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
10. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109–113 (2012).
11. Kim, T. K. et al. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010).
12. Sebe-Pedros, A. et al. The dynamic regulatory genome of capsaspora and the origin of animal multicellularity. *Cell* **165**, 1224–1237 (2016).
13. Weber, B., Zicola, J., Oka, R. & Stam, M. Plant enhancers: a call for discovery. *Trends Plant Sci.* **21**, 974–987 (2016).
14. Marand, A. P., Zhang, T., Zhu, B. & Jiang, J. M. Towards genome-wide prediction and characterization of enhancers in plants. *Biochim. Biophys. Acta* **1860**, 131–139 (2017).
15. Salvi, S. et al. Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. *Proc. Natl Acad. Sci. USA* **104**, 11376–11381 (2007).
16. Louwers, M. et al. Tissue- and expression level-specific chromatin looping at maize b1 epialleles. *Plant Cell* **21**, 832–842 (2009).
17. Xu, G. et al. Complex genetic architecture underlies maize tassel domestication. *New Phytol.* **214**, 852–864 (2017).
18. Studer, A., Zhao, Q., Ross-Ibarra, J. & Doebley, J. Identification of a functional transposon insertion in the maize domestication gene tb1. *Nat. Genet.* **43**, 1160–1163 (2011).
19. Adrian, J. et al. *cis*-Regulatory elements and chromatin state coordinately control temporal and spatial expression of *FLOWERING LOCUS T* in *Arabidopsis*. *Plant Cell* **22**, 1425–1440 (2010).
20. McGarry, R. C. & Ayre, B. G. A. DNA element between At4g28630 and At4g28640 confers companion-cell specific expression following the sink-to-source transition in mature minor vein phloem. *Planta* **228**, 839–849 (2008).
21. Yang, W. et al. An egg apparatus-specific enhancer of *Arabidopsis*, identified by enhancer detection. *Plant Physiol.* **139**, 1421–1432 (2005).
22. Liu, L. et al. Induced and natural variation of promoter length modulates the photoperiodic response of *FLOWERING LOCUS T*. *Nat. Commun.* **5**, 4558 (2014).
23. Rodgers-Melnick, E., Vera, D. L., Bass, H. W. & Buckler, E. S. Open chromatin reveals the functional maize genome. *Proc. Natl Acad. Sci. USA* **113**, E3177–E3184 (2016).
24. Dong, P. et al. 3D chromatin architecture of large plant genomes determined by local A/B compartments. *Mol. Plant* **10**, 1497–1509 (2017).
25. Li, X. et al. High-resolution mapping of epigenetic modifications of the rice genome uncovers interplay between DNA methylation, histone methylation, and gene expression. *Plant Cell* **20**, 259–276 (2008).
26. Qiu, Z. K. et al. Identification of regulatory DNA elements using genome-wide mapping of DNase I hypersensitive sites during tomato fruit development. *Mol. Plant* **9**, 1168–1182 (2016).
27. Zhang, W. et al. High-resolution mapping of open chromatin in the rice genome. *Genome Res.* **22**, 151–162 (2012).
28. Zhang, W., Zhang, T., Wu, Y. & Jiang, J. Open chromatin in plant genomes. *Cytogenet. Genome Res.* **143**, 18–27 (2014).
29. Zhang, W. L., Zhang, T., Wu, Y. F. & Jiang, J. M. Genome-wide identification of regulatory DNA elements and protein-binding footprints using signatures of open chromatin in *Arabidopsis*. *Plant Cell* **24**, 2719–2731 (2012).
30. Lu, P. et al. Genome encode analyses reveal the basis of convergent evolution of fleshy fruit ripening. *Nat. Plants* **4**, 784–791 (2018).
31. Chua, Y. L., Watson, L. A. & Gray, J. C. The transcriptional enhancer of the pea plastocyanin gene associates with the nuclear matrix and regulates gene expression through histone acetylation. *Plant Cell* **15**, 1468–1479 (2003).
32. Oka, R. et al. Genome-wide mapping of transcriptional enhancer candidates using DNA and chromatin features in maize. *Genome Biol.* **18**, 137 (2017).
33. Lu, Z., Ricci, W. A., Schmitz, R. J. & Zhang, X. Identification of *cis*-regulatory elements by chromatin structure. *Curr. Opin. Plant Biol.* **42**, 90–94 (2018).
34. Tsompana, M. & Buck, M. J. Chromatin accessibility: a window into the genome. *Epigenet. Chromatin* **7**, 33 (2014).
35. Zhu, B., Zhang, W. L., Zhang, T., Liu, B. & Jiang, J. M. Genome-wide prediction and validation of intergenic enhancers in *Arabidopsis* using open chromatin signatures. *Plant Cell* **27**, 2415–2426 (2015).
36. Sullivan, A. M. et al. Mapping and dynamics of regulatory DNA and transcription factor networks in *A. thaliana*. *Cell Rep.* **8**, 2015–2030 (2014).
37. Sijacic, P., Bajic, M., McKinney, E. C., Meagher, R. B. & Deal, R. B. Changes in chromatin accessibility between Arabidopsis stem cells and mesophyll cells illuminate cell type-specific transcription factor networks. *Plant J.* **94**, 215–231 (2018).
38. O'Malley, R. C. et al. Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell* **165**, 1280–1292 (2016).
39. 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. *Plant Cell* **30**, 15–36 (2017).
40. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **109**, 1–9 (2015).
41. Lu, Z., Hofmeister, B. T., Vollmers, C., DuBois, R. M. & Schmitz, R. J. Combining ATAC-seq with nuclei sorting for discovery of *cis*-regulatory regions in plant genomes. *Nucleic Acids Res.* **45**, e41 (2017).
42. Bajic, M., Maher, K. A. & Deal, R. B. Identification of open chromatin regions in plant genomes using ATAC-seq. *Methods Mol. Biol.* **1675**, 183–201 (2018).
43. Haring, M. et al. The role of DNA methylation, nucleosome occupancy and histone modifications in paramutation. *Plant J.* **63**, 366–378 (2010).
44. Wang, J. et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* **22**, 1798–1812 (2012).
45. Grigoriev, I. V. et al. The Genome Portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Res.* **40**, D26–D32 (2012).
46. Nordberg, H. et al. The Genome Portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Res.* **42**, D26–D31 (2014).
47. Turco, G., Schnable, J. C., Pedersen, B. & Freeling, M. Automated conserved non-coding sequence (CNS) discovery reveals differences in gene content and promoter evolution among grasses. *Front. Plant Sci.* **4**, 170 (2013).
48. Kidwell, M. G. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* **115**, 49–63 (2002).
49. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86 (2017).
50. Bejerano, G. et al. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441**, 87–90 (2006).
51. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**, 1083–1087 (2016).
52. Flemr, M. et al. A retrotransposon-driven dicer isoform directs endogenous small interfering RNA production in mouse oocytes. *Cell* **155**, 807–816 (2013).
53. Moriyama, Y. & Koshiba-Takeuchi, K. Significance of whole-genome duplications on the emergence of evolutionary novelties. *Brief. Funct. Genom.* **17**, 329–338 (2018).
54. Schnable, J. C., Springer, N. M. & Freeling, M. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl Acad. Sci. USA* **108**, 4069–4074 (2011).
55. Schmutz, J. et al. Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010).
56. Rada-Iglesias, A. et al. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–283 (2011).
57. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res.* **21**, 381–395 (2011).
58. Creyghton, M. P. et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl Acad. Sci. USA* **107**, 21931–21936 (2010).
59. Zhang, X. et al. Whole-genome analysis of histone H3 lysine 27 trimethylation in *Arabidopsis*. *PLoS Biol.* **5**, e129 (2007).
60. Wiles, E. T. & Selker, E. U. H3K27 methylation: a promiscuous repressive chromatin mark. *Curr. Opin. Genet. Dev.* **43**, 31–37 (2017).
61. Ricci, W. A. et al. Widespread long-range *cis*-regulatory elements in the maize genome. *Nat. Plants* https://doi.org/10.1038/s41477-019-0547-0 (2019).
62. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
63. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

64. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
65. Harper, L., Gardiner, J., Andorf, C. & Lawrence, C. J. MaizeGDB: the maize genetics and genomics database. *Methods Mol. Biol.* **1374**, 187–202 (2016).
66. Jiao, Y. et al. Improved maize reference genome with single-molecule technologies. *Nature* **546**, 524–527 (2017).
67. Mascher, M. et al. A chromosome conformation capture ordered sequence of the barley genome. *Nature* **544**, 427–433 (2017).
68. Harkess, A. et al. The asparagus genome sheds light on the origin and evolution of a young Y chromosome. *Nat. Commun.* **8**, 1279 (2017).
69. Michael, T. P. et al. Comprehensive definition of genome features in *Spirodela polyrhiza* by high-depth physical mapping and short-read DNA sequencing strategies. *Plant J.* **89**, 617–635 (2017).
70. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
71. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-seq. *Bioinformatics* **25**, 1105–1111 (2009).
72. Trapnell, C. et al. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
73. Vera Alvarez, R., Pongor, L. S., Marino-Ramirez, L. & Landsman, D. TPMCalculator: one-step software to quantify mRNA abundance of genomic features. *Bioinformatics* **35**, 1960–1962 (2018).
74. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
75. Zhang, Y. et al. Model-based analysis of ChIP-seq (MACS). *Genome Biol.* **9**, R137 (2008).
76. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
77. Wang, Y. P. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
78. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
79. Niederhuth, C. E. et al. Widespread natural variation of DNA methylation within angiosperms. *Genome Biol.* **17**, 194 (2016).
80. Ramirez, F., Dundar, F., Diehl, S., Gruning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).
81. Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. Nbclust: an R package for determining the relevant number of clusters in a data set. *J. Stat. Softw.* **61**, 1–36 (2014).
82. RepeatMasker v.Open-4.0 (Smit, A.F.A., Hubley, R. & Green, P.; 2013–2015); http://www.repeatmasker.org

## Acknowledgements

We thank R. Deal for providing the H2A.Z antibodies used in this study. This work was funded by the NSF IOS-1546867 and NSF IOS-1856627 to R.J.S. and X.Z., NSF IOS-1238142 to X.Z. and NSF IOS-1339194 to R.J.S. R.J.S. acknowledges support from the Technical University of Munich-Institute for Advanced Study funded by the German Excellent Initiative and the European Seventh Framework Programme under grant agreement no. 291763. R.J.S. is a Pew Scholar in the Biomedical Sciences, supported by The Pew Charitable Trusts.

## Author contributions

X.Z. and R.J.S conceived and designed experiments. Z.L. performed experiments. Z.L., A.P.M., W.A.R. and C.L.E. analysed the data. Z.L., X.Z. and R.J.S. wrote the paper.

## Competing interests

R.J.S. and X.Z. are co-founders of REquest Genomics, LLC, a company that provides epigenomics services.

## Additional information

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41477-019-0548-z.

**Correspondence and requests for materials** should be addressed to X.Z. or R.J.S.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Robert J Schmitz & Xiaoyu Zhang

Last updated by author(s): 05/15/2019

# Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed |
|---|---|
| ☐ | ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☒ | ☐ A description of all covariates tested |
| ☐ | ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☒ | ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☐ | ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☒ | ☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | No software was used for data collection. |
| Data analysis | Trimmomatic (v0.36);Bowtie (v1.1.1); Bowtie2(v2.1.1); Picard (v2.16.0); TOPHAT (v2.1.1); Cufflink (v2.2.1); TPMCalculator; MACS2(v2.1.2); BEDTools (v2.26.0); CNS Discovery Pipeline (v3.0); Multiple Collinearity Scan toolkit (MCscanX); BLASTN (v2.28.0+); deepTools (v3.0.2); RepeatMasker (v4.07); samtools (v1.3.1); blastn (v2.2.29+); R (v3.4.3); RStudio (v1.1.383). |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data generated from this study has been uploaded to the Gene Expression Omnibus (GEO) database and can be retrieved through accession number GSE128434.

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

# Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Sample size | For ATAC-seq, 200mg freshly collected or flash frozen tissues from a collection of 10-50 individual leaves/seedlings were used; for ChIP-seq and RNA-seq, 1g of tissues were used. No statistical methods were used to predetermine sample sizes. |
| Data exclusions | No data exclusions. |
| Replication | All data presented in figure legends and methods. ATAC-Seq data were replicated twice; ChIP-seq data were replicated once and RNA-seq were replicated twice. Replications were successfully implemented. |
| Randomization | For each species, the plants used for ATAC-seq, ChIP-seq and RNA-seq were all grown in the same conditions (see Methods) and were placed randomly in the growth room. |
| Blinding | For each experiment, plants were randomly selected. Blinding was not performed for sample collection and/or data analyses as the species used are quite diverse in phenotype and the patterns of chromatin modifications are unique. |

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☐ | ☒ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology |
| ☒ | ☐ Animals and other organisms |
| ☒ | ☐ Human research participants |
| ☒ | ☐ Clinical data |

## Methods

| n/a | Involved in the study |
|---|---|
| ☐ | ☒ ChIP-seq |
| ☐ | ☒ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |

## Antibodies

| | |
|---|---|
| Antibodies used | H3: Abcam, cat # ab1791, lot # GR71822-1; H3K4me1: Abcam, cat # ab8895, lot # 889421; H3K4me3: Millipore, cat # 07-473, lot # 28; 9113; H3K27me3: Millipore, cat # 07-449, lot # DAM1703508; H3K36me3: Abcam, cat # ab9050, lot # 826243; H3K56ac: Millipore, cat # 07-677-1, lot # 2514206; anti-H2A.Z: provided by Dr. Roger Deal |
| Validation | All antibodies used here have been validated by manufacturers. Furthermore, these antibodies have a long historical use in the ENCODE project and in plant genomics. H3K56ac and H3K27me3 have been independently validated with peptide array. We have also validated antibodies with computational analysis. The histone marks H3K56ac, H3K27me3, H3K36me3, and H3K4me3 were used for cluster analysis. Gene metaplot analysis demonstrates that these marks have distinct enrichment profiles, indicating that cross-reactions were not problematic in our methods. The antibody provided by Dr. Roger Deal was validated in a previous study: E. Shannon Torres and Roger B. Deal, The histone variant H2A.Z and chromatin remodeler BRAHMA act coordinately and antagonistically to regulate transcription and nucleosome dynamics in Arabidopsis. Plant J. (2019). |

## ChIP-seq

### Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
*May remain private before publication.*

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE128434 (token: gzunomwqxlsxzoh)

Files in database submission

ChIP_Arabidopsis_7days_leaf_H2A.Z.bw;
ChIP_Arabidopsis_7days_leaf_H2A.Z.fastq.gz;
ChIP_Arabidopsis_7days_leaf_H3.bw;
ChIP_Arabidopsis_7days_leaf_H3.fastq.gz;
ChIP_Arabidopsis_7days_leaf_H3K27me3.bw;
ChIP_Arabidopsis_7days_leaf_H3K27me3.fastq.gz;
ChIP_Arabidopsis_7days_leaf_H3K36me3.bw;
ChIP_Arabidopsis_7days_leaf_H3K36me3.fastq.gz;
ChIP_Arabidopsis_7days_leaf_H3K4me1.bw;
ChIP_Arabidopsis_7days_leaf_H3K4me1.fastq.gz;
ChIP_Arabidopsis_7days_leaf_H3K4me3.bw;
ChIP_Arabidopsis_7days_leaf_H3K4me3.fastq.gz;
ChIP_Arabidopsis_7days_leaf_H3K56ac.bw;
ChIP_Arabidopsis_7days_leaf_H3K56ac.fastq.gz;
ChIP_Arabidopsis_7days_leaf_Input.bw;
ChIP_Arabidopsis_7days_leaf_Input.fastq.gz;
ChIP_Asparagus_10days_leaf_H2A.Z.bw;
ChIP_Asparagus_10days_leaf_H2A.Z.fastq.gz;
ChIP_Asparagus_10days_leaf_H3.bw;
ChIP_Asparagus_10days_leaf_H3.fastq.gz;
ChIP_Asparagus_10days_leaf_H3K27me3.bw;
ChIP_Asparagus_10days_leaf_H3K27me3.fastq.gz;
ChIP_Asparagus_10days_leaf_H3K36me3.bw;
ChIP_Asparagus_10days_leaf_H3K36me3.fastq.gz;
ChIP_Asparagus_10days_leaf_H3K4me1.bw;
ChIP_Asparagus_10days_leaf_H3K4me1.fastq.gz;
ChIP_Asparagus_10days_leaf_H3K4me3.bw;
ChIP_Asparagus_10days_leaf_H3K4me3.fastq.gz;
ChIP_Asparagus_10days_leaf_H3K56ac.bw;
ChIP_Asparagus_10days_leaf_H3K56ac.fastq.gz;
ChIP_Asparagus_10days_leaf_Input.bw;
ChIP_Asparagus_10days_leaf_Input.fastq.gz;
ChIP_Barley_7days_leaf_H2A.Z.bw;
ChIP_Barley_7days_leaf_H2A.Z.fastq.gz;
ChIP_Barley_7days_leaf_H3.bw;
ChIP_Barley_7days_leaf_H3.fastq.gz;
ChIP_Barley_7days_leaf_H3K27me3.bw;
ChIP_Barley_7days_leaf_H3K27me3.fastq.gz;
ChIP_Barley_7days_leaf_H3K36me3.bw;
ChIP_Barley_7days_leaf_H3K36me3.fastq.gz;
ChIP_Barley_7days_leaf_H3K4me1.bw;
ChIP_Barley_7days_leaf_H3K4me1.fastq.gz;
ChIP_Barley_7days_leaf_H3K4me3.bw;
ChIP_Barley_7days_leaf_H3K4me3.fastq.gz;
ChIP_Barley_7days_leaf_H3K56ac.bw;
ChIP_Barley_7days_leaf_H3K56ac.fastq.gz;
ChIP_Barley_7days_leaf_Input.bw;
ChIP_Barley_7days_leaf_Input.fastq.gz;
ChIP_Brachypodium_7days_leaf_H2A.Z.bw;
ChIP_Brachypodium_7days_leaf_H2A.Z.fastq.gz;
ChIP_Brachypodium_7days_leaf_H3.bw;
ChIP_Brachypodium_7days_leaf_H3.fastq.gz;
ChIP_Brachypodium_7days_leaf_H3K27me3.bw;
ChIP_Brachypodium_7days_leaf_H3K27me3.fastq.gz;
ChIP_Brachypodium_7days_leaf_H3K36me3.bw;
ChIP_Brachypodium_7days_leaf_H3K36me3.fastq.gz;
ChIP_Brachypodium_7days_leaf_H3K4me1.bw;
ChIP_Brachypodium_7days_leaf_H3K4me1.fastq.gz;
ChIP_Brachypodium_7days_leaf_H3K4me3.bw;
ChIP_Brachypodium_7days_leaf_H3K4me3.fastq.gz;
ChIP_Brachypodium_7days_leaf_H3K56ac.bw;
ChIP_Brachypodium_7days_leaf_H3K56ac.fastq.gz;
ChIP_Brachypodium_7days_leaf_Input.bw;
ChIP_Brachypodium_7days_leaf_Input.fastq.gz;
ChIP_duckweed_3days_after_tranferring_H2A.Z.bw;
ChIP_duckweed_3days_after_tranferring_H2A.Z.fastq.gz;
ChIP_duckweed_3days_after_tranferring_H3.bw;
ChIP_duckweed_3days_after_tranferring_H3.fastq.gz;
ChIP_duckweed_3days_after_tranferring_H3K27me3.bw;
ChIP_duckweed_3days_after_tranferring_H3K27me3.fastq.gz;
ChIP_duckweed_3days_after_tranferring_H3K36me3.bw;

ChIP_duckweed_3days_after_tranferring_H3K36me3.fastq.gz;
ChIP_duckweed_3days_after_tranferring_H3K4me1.bw;
ChIP_duckweed_3days_after_tranferring_H3K4me1.fastq.gz;
ChIP_duckweed_3days_after_tranferring_H3K4me3.bw;
ChIP_duckweed_3days_after_tranferring_H3K4me3.fastq.gz;
ChIP_duckweed_3days_after_tranferring_H3K56ac.bw;
ChIP_duckweed_3days_after_tranferring_H3K56ac.fastq.gz;
ChIP_duckweed_3days_after_tranferring_Input.bw;
ChIP_duckweed_3days_after_tranferring_Input.fastq.gz;
ChIP_Eutrema_10days_leaf_H2A.Z.bw;
ChIP_Eutrema_10days_leaf_H2A.Z.fastq.gz;
ChIP_Eutrema_10days_leaf_H3.bw;
ChIP_Eutrema_10days_leaf_H3.fastq.gz;
ChIP_Eutrema_10days_leaf_H3K27me3.bw;
ChIP_Eutrema_10days_leaf_H3K27me3.fastq.gz;
ChIP_Eutrema_10days_leaf_H3K36me3.bw;
ChIP_Eutrema_10days_leaf_H3K36me3.fastq.gz;
ChIP_Eutrema_10days_leaf_H3K4me1.bw;
ChIP_Eutrema_10days_leaf_H3K4me1.fastq.gz;
ChIP_Eutrema_10days_leaf_H3K4me3.bw;
ChIP_Eutrema_10days_leaf_H3K4me3.fastq.gz;
ChIP_Eutrema_10days_leaf_H3K56ac.bw;
ChIP_Eutrema_10days_leaf_H3K56ac.fastq.gz;
ChIP_Eutrema_10days_leaf_Input.bw;
ChIP_Eutrema_10days_leaf_Input.fastq.gz;
ChIP_Maize_7days_leaf_H2A.Z.bw;
ChIP_Maize_7days_leaf_H2A.Z.fastq.gz;
ChIP_Maize_7days_leaf_H3.bw;
ChIP_Maize_7days_leaf_H3.fastq.gz;
ChIP_Maize_7days_leaf_H3K27me3.bw;
ChIP_Maize_7days_leaf_H3K27me3.fastq.gz;
ChIP_Maize_7days_leaf_H3K36me3.bw;
ChIP_Maize_7days_leaf_H3K36me3.fastq.gz;
ChIP_Maize_7days_leaf_H3K4me1.bw;
ChIP_Maize_7days_leaf_H3K4me1.fastq.gz;
ChIP_Maize_7days_leaf_H3K4me3.bw;
ChIP_Maize_7days_leaf_H3K4me3.fastq.gz;
ChIP_Maize_7days_leaf_H3K56ac.bw;
ChIP_Maize_7days_leaf_H3K56ac.fastq.gz;
ChIP_Maize_7days_leaf_Input.bw;
ChIP_Maize_7days_leaf_Input.fastq.gz;
ChIP_Phaseolus_10days_leaf_H2A.Z.bw;
ChIP_Phaseolus_10days_leaf_H2A.Z.fastq.gz;
ChIP_Phaseolus_10days_leaf_H3.bw;
ChIP_Phaseolus_10days_leaf_H3.fastq.gz;
ChIP_Phaseolus_10days_leaf_H3K27me3.bw;
ChIP_Phaseolus_10days_leaf_H3K27me3.fastq.gz;
ChIP_Phaseolus_10days_leaf_H3K36me3.bw;
ChIP_Phaseolus_10days_leaf_H3K36me3.fastq.gz;
ChIP_Phaseolus_10days_leaf_H3K4me1.bw;
ChIP_Phaseolus_10days_leaf_H3K4me1.fastq.gz;
ChIP_Phaseolus_10days_leaf_H3K4me3.bw;
ChIP_Phaseolus_10days_leaf_H3K4me3.fastq.gz;
ChIP_Phaseolus_10days_leaf_H3K56ac.bw;
ChIP_Phaseolus_10days_leaf_H3K56ac.fastq.gz;
ChIP_Phaseolus_10days_leaf_Input.bw;
ChIP_Phaseolus_10days_leaf_Input.fastq.gz;
ChIP_Populus_leaf_H2A.Z.bw;
ChIP_Populus_leaf_H2A.Z.fastq.gz;
ChIP_Populus_leaf_H3.bw;
ChIP_Populus_leaf_H3.fastq.gz;
ChIP_Populus_leaf_H3K27me3.bw;
ChIP_Populus_leaf_H3K27me3.fastq.gz;
ChIP_Populus_leaf_H3K36me3.bw;
ChIP_Populus_leaf_H3K36me3.fastq.gz;
ChIP_Populus_leaf_H3K4me1.bw;
ChIP_Populus_leaf_H3K4me1.fastq.gz;
ChIP_Populus_leaf_H3K4me3.bw;
ChIP_Populus_leaf_H3K4me3.fastq.gz;
ChIP_Populus_leaf_H3K56ac.bw;
ChIP_Populus_leaf_H3K56ac.fastq.gz;
ChIP_Populus_leaf_Input.bw;
ChIP_Populus_leaf_Input.fastq.gz;
ChIP_Rice_7days_leaf_H2A.Z.bw;
ChIP_Rice_7days_leaf_H2A.Z.fastq.gz;

ChIP_Rice_7days_leaf_H3.bw;
ChIP_Rice_7days_leaf_H3.fastq.gz;
ChIP_Rice_7days_leaf_H3K27me3.bw;
ChIP_Rice_7days_leaf_H3K27me3.fastq.gz;
ChIP_Rice_7days_leaf_H3K36me3.bw;
ChIP_Rice_7days_leaf_H3K36me3.fastq.gz;
ChIP_Rice_7days_leaf_H3K4me1.bw;
ChIP_Rice_7days_leaf_H3K4me1.fastq.gz;
ChIP_Rice_7days_leaf_H3K4me3.bw;
ChIP_Rice_7days_leaf_H3K4me3.fastq.gz;
ChIP_Rice_7days_leaf_H3K56ac.bw;
ChIP_Rice_7days_leaf_H3K56ac.fastq.gz;
ChIP_Rice_7days_leaf_Input.bw;
ChIP_Rice_7days_leaf_Input.fastq.gz;
ChIP_Setaria_7days_leaf_H2A.Z.bw;
ChIP_Setaria_7days_leaf_H2A.Z.fastq.gz;
ChIP_Setaria_7days_leaf_H3.bw;
ChIP_Setaria_7days_leaf_H3.fastq.gz;
ChIP_Setaria_7days_leaf_H3K27me3.bw;
ChIP_Setaria_7days_leaf_H3K27me3.fastq.gz;
ChIP_Setaria_7days_leaf_H3K36me3.bw;
ChIP_Setaria_7days_leaf_H3K36me3.fastq.gz;
ChIP_Setaria_7days_leaf_H3K4me1.bw;
ChIP_Setaria_7days_leaf_H3K4me1.fastq.gz;
ChIP_Setaria_7days_leaf_H3K4me3.bw;
ChIP_Setaria_7days_leaf_H3K4me3.fastq.gz;
ChIP_Setaria_7days_leaf_H3K56ac.bw;
ChIP_Setaria_7days_leaf_H3K56ac.fastq.gz;
ChIP_Setaria_7days_leaf_Input.bw;
ChIP_Setaria_7days_leaf_Input.fastq.gz;
ChIP_Sorghum_7days_leaf_H2A.Z.bw;
ChIP_Sorghum_7days_leaf_H2A.Z.fastq.gz;
ChIP_Sorghum_7days_leaf_H3.bw;
ChIP_Sorghum_7days_leaf_H3.fastq.gz;
ChIP_Sorghum_7days_leaf_H3K27me3.bw;
ChIP_Sorghum_7days_leaf_H3K27me3.fastq.gz;
ChIP_Sorghum_7days_leaf_H3K36me3.bw;
ChIP_Sorghum_7days_leaf_H3K36me3.fastq.gz;
ChIP_Sorghum_7days_leaf_H3K4me1.bw;
ChIP_Sorghum_7days_leaf_H3K4me1.fastq.gz;
ChIP_Sorghum_7days_leaf_H3K4me3.bw;
ChIP_Sorghum_7days_leaf_H3K4me3.fastq.gz;
ChIP_Sorghum_7days_leaf_H3K56ac.bw;
ChIP_Sorghum_7days_leaf_H3K56ac.fastq.gz;
ChIP_Sorghum_7days_leaf_Input.bw;
ChIP_Sorghum_7days_leaf_Input.fastq.gz;
ChIP_Soybean_10days_leaf_H2A.Z.bw;
ChIP_Soybean_10days_leaf_H2A.Z.fastq.gz;
ChIP_Soybean_10days_leaf_H3.bw;
ChIP_Soybean_10days_leaf_H3.fastq.gz;
ChIP_Soybean_10days_leaf_H3K27me3.bw;
ChIP_Soybean_10days_leaf_H3K27me3.fastq.gz;
ChIP_Soybean_10days_leaf_H3K36me3.bw;
ChIP_Soybean_10days_leaf_H3K36me3.fastq.gz;
ChIP_Soybean_10days_leaf_H3K4me1.bw;
ChIP_Soybean_10days_leaf_H3K4me1.fastq.gz;
ChIP_Soybean_10days_leaf_H3K4me3.bw;
ChIP_Soybean_10days_leaf_H3K4me3.fastq.gz;
ChIP_Soybean_10days_leaf_H3K56ac.bw;
ChIP_Soybean_10days_leaf_H3K56ac.fastq.gz;
ChIP_Soybean_10days_leaf_Input.bw;
ChIP_Soybean_10days_leaf_Input.fastq.gz.

Genome browser session
(e.g. UCSC)

http://epigenome.genetics.uga.edu/PlantEpigenome/

## Methodology

Replicates

Approximately 1g of freshly collected leaves or flash frozen leaves from a collection of 10-50 individual seedlings/leaves were used for the assays. One replicate was performed for each histone modfiication.

Sequencing depth

The sequence depth and mapping statistics are shown in Supplementary Tables 15.

Antibodies

anti-H3K4me1 : Abcam, cat# ab8895; anti-H3K4me3 : Millipore -sigma, cat# 07-473; anti-H3 : Abcam, cat# ab1791; anti-H3K36me3 : Abcam, cat# ab9050; anti-H3K27me3 : Millipore -sigma, cat# 07-449; anti-H3K56ac : Millipore-sigma, cat# 07-667; anti-H2A.Z : from Roger Deal's lab.

Peak calling parameters

The peaks for ChIP-seq of H2A.Z, H3K27me3, H3K36me3 and H3K4me1 were called by MACS2 with "--nomodel --extsize 147 --broad --broad-cutoff 0.1" and FDR<0.05; for H3K4me3 and H3K56ac were called by MACS2 with "--nomodel --extsize 147" and FDR < 0.05.

Data quality

Peak counts identified by MACS.
Asparagus_10days_leaf-H2A.Z : 75,696;
Asparagus_10days_leaf-H3K27me3 : 20,775;
Asparagus_10days_leaf-H3K36me3 : 47,293;
Asparagus_10days_leaf-H3K4me3 : 40,908;
Asparagus_10days_leaf-H3K56ac : 28,393;
Asparagus_10days_leaf-H3K4me1 : 89,711;
Arabidopsis_7days_leaf-H2A.Z : 31,484;
Arabidopsis_7days_leaf-H3K27me3 : 9,893;
Arabidopsis_7days_leaf-H3K36me3 : 14,685;
Arabidopsis_7days_leaf-H3K4me3 : 17,812;
Arabidopsis_7days_leaf-H3K56ac : 19,504;
Arabidopsis_7days_leaf-H3K4me1 : 23,944;
Brachypodium_7days_leaf-H2A.Z : 55,255;
Brachypodium_7days_leaf-H3K27me3 : 16,356;
Brachypodium_7days_leaf-H3K36me3 : 30,825;
Brachypodium_7days_leaf-H3K4me3 : 40,550;
Brachypodium_7days_leaf-H3K56ac : 35,102;
Brachypodium_7days_leaf-H3K4me1 : 51,118;
Eutrema_10days_leaf-H2A.Z : 30,265;
Eutrema_10days_leaf-H3K27me3 : 15,876;
Eutrema_10days_leaf-H3K36me3 : 19,263;
Eutrema_10days_leaf-H3K4me3 : 24,465;
Eutrema_10days_leaf-H3K56ac : 25,341;
Eutrema_10days_leaf-H3K4me1 : 30,378;
Soybean_10days_leaf-H2A.Z : 120,191;
Soybean_10days_leaf-H3K27me3 : 37,575;
Soybean_10days_leaf-H3K36me3 : 50,280;
Soybean_10days_leaf-H3K4me3 : 61,986;
Soybean_10days_leaf-H3K56ac : 61,730;
Soybean_10days_leaf-H3K4me1 : 56,042;
Barley_7days_leaf-H2A.Z : 95,370;
Barley_7days_leaf-H3K27me3 : 55,291;
Barley_7days_leaf-H3K36me3 : 27,778;
Barley_7days_leaf-H3K4me3 : 53,680;
Barley_7days_leaf-H3K56ac : 33,746;
Barley_7days_leaf-H3K4me1 : 171,164;
Rice_7days_leaf-H2A.Z : 68,871;
Rice_7days_leaf-H3K27me3 : 22,775;
Rice_7days_leaf-H3K36me3 : 27,899;
Rice_7days_leaf-H3K4me3 : 25,282;
Rice_7days_leaf-H3K56ac : 23,104;
Rice_7days_leaf-H3K4me1 : 42,808;
Populus_leaf-H2A.Z : 59,107;
Populus_leaf-H3K27me3 : 29,795;
Populus_leaf-H3K36me3 : 29,311;
Populus_leaf-H3K4me3 : 23,286;
Populus_leaf-H3K56ac : 13,314;
Populus_leaf-H3K4me1 : 51,806;
Phaseolus_10days_leaf-H2A.Z : 61,673;
Phaseolus_10days_leaf-H3K27me3 : 22,183;
Phaseolus_10days_leaf-H3K36me3 : 22,452;
Phaseolus_10days_leaf-H3K4me3 : 19,942;
Phaseolus_10days_leaf-H3K56ac : 15,332;
Phaseolus_10days_leaf-H3K4me1 : 55,594;
Sorghum_7days_leaf-H2A.Z : 67,059;
Sorghum_7days_leaf-H3K27me3 : 18,696;
Sorghum_7days_leaf-H3K36me3 : 20,730;
Sorghum_7days_leaf-H3K4me3 : 25,764;
Sorghum_7days_leaf-H3K56ac : 20,438;
Sorghum_7days_leaf-H3K4me1 : 20,091;
duckweed_3days_after_tranferring-H2A.Z : 17,295;
duckweed_3days_after_tranferring-H3K27me3 : 8,903;
duckweed_3days_after_tranferring-H3K36me3 : 19,237;
duckweed_3days_after_tranferring-H3K4me3 : 24,069;
duckweed_3days_after_tranferring-H3K56ac : 22,018;

duckweed_3days_after_tranferring-H3K4me1 : 23,208;
Setaria_7days_leaf-H2A.Z : 67,973;
Setaria_7days_leaf-H3K27me3 : 23,197;
Setaria_7days_leaf-H3K36me3 : 18,871;
Setaria_7days_leaf-H3K4me3 : 40,443;
Setaria_7days_leaf-H3K56ac : 7,969;
Setaria_7days_leaf-H3K4me1 : 22,294;
Maize_7days_leaf-H2A.Z : 84,546;
Maize_7days_leaf-H3K27me3 : 18,059;
Maize_7days_leaf-H3K36me3 : 41,134;
Maize_7days_leaf-H3K4me3 : 50,094;
Maize_7days_leaf-H3K56ac : 60,628;
Maize_7days_leaf-H3K4me1 : 56,992.

Software

Trimmomatic (v0.36);Bowtie (v1.1.1); Bowtie2(v2.1.1); Picard (v2.16.0); deepTools (v3.0.2); BEDTools (v2.26.0); samtools (v1.3.1); MACS2 (v2.1.2).

# Flow Cytometry

## Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

## Methodology

Sample preparation

Approximately 200 mg of freshly collected leaves or flash frozen leaves from a collection of 10-50 individual seedlings were immediately chopped with a razor blade in ~ 1ml of pre-chilled lysis buffer (15 mM Tris-HCl pH 7.5, 20 mM NaCl, 80 mM KCl, 0.5 mM spermine, 5 mM 2-Mercaptoethanol, 0.2% TritonX-100). The chopped slurry was filtered twice through miracloth and once through a 40 µm filter. The crude nuclei were stained with DAPI and loaded into a flow cytometer.

Instrument

Beckman Coulter MoFlo XDP and Beckman Coulter MoFlo Astrios EQ.

Software

Summit version 6.3.1 for the XDP and version 6.3.1 for Astrios.

Cell population abundance

For each library preparation, 50,000 nuclei were used.

Gating strategy

There are multiple DAPI signal peaks with high quality nuclei, reflecting the copy number of plant genomes. The nuclei with DAPI signal >= that of 2×n genomes were collected for the ATAC-seq.

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.