nature plants

SUPPLEMENTARY INFORMATION

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

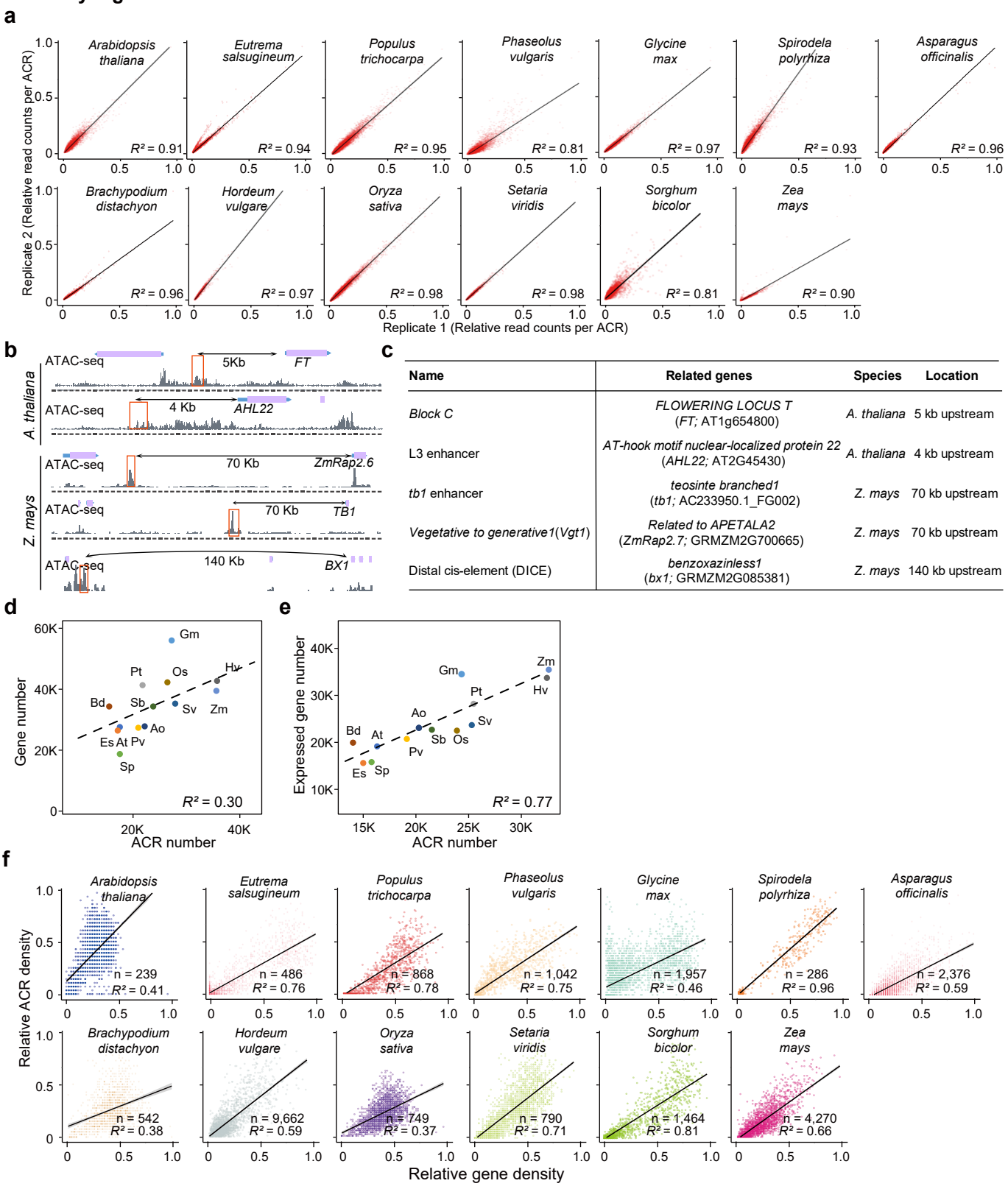
In the format provided by the authors and unedited.

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*]

[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

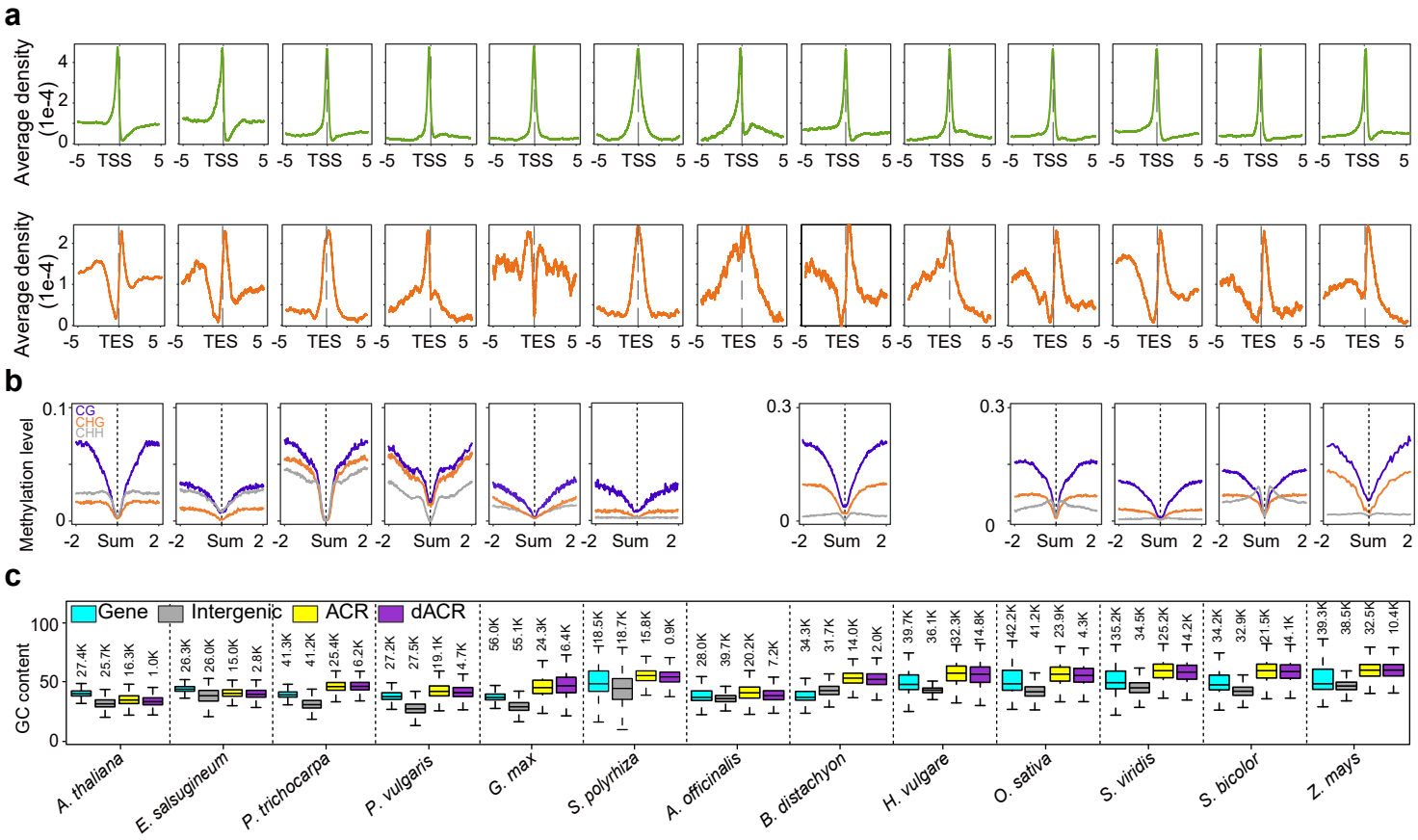
Supplementary Fig. 1



Name	Related genes	Species	Location
Block C	*FLOWERING LOCUS T* (*FT;* AT1g654800)	*A. thaliana*	5 kb upstream
L3 enhancer	*AT-hook motif nuclear-localized protein 22* (*AHL22;* AT2G45430)	*A. thaliana*	4 kb upstream
tb1 enhancer	*teosinte branched1* (*tb1;* AC233950.1_FG002)	*Z. mays*	70 kb upstream
Vegetative to generative1(*Vgt1*)	*Related to APETALA2* (*ZmRap2.7;* GRMZM2G700665)	*Z. mays*	70 kb upstream
Distal cis-element (DICE)	*benzoxazinless1* (*bx1;* GRMZM2G085381)	*Z. mays*	140 kb upstream

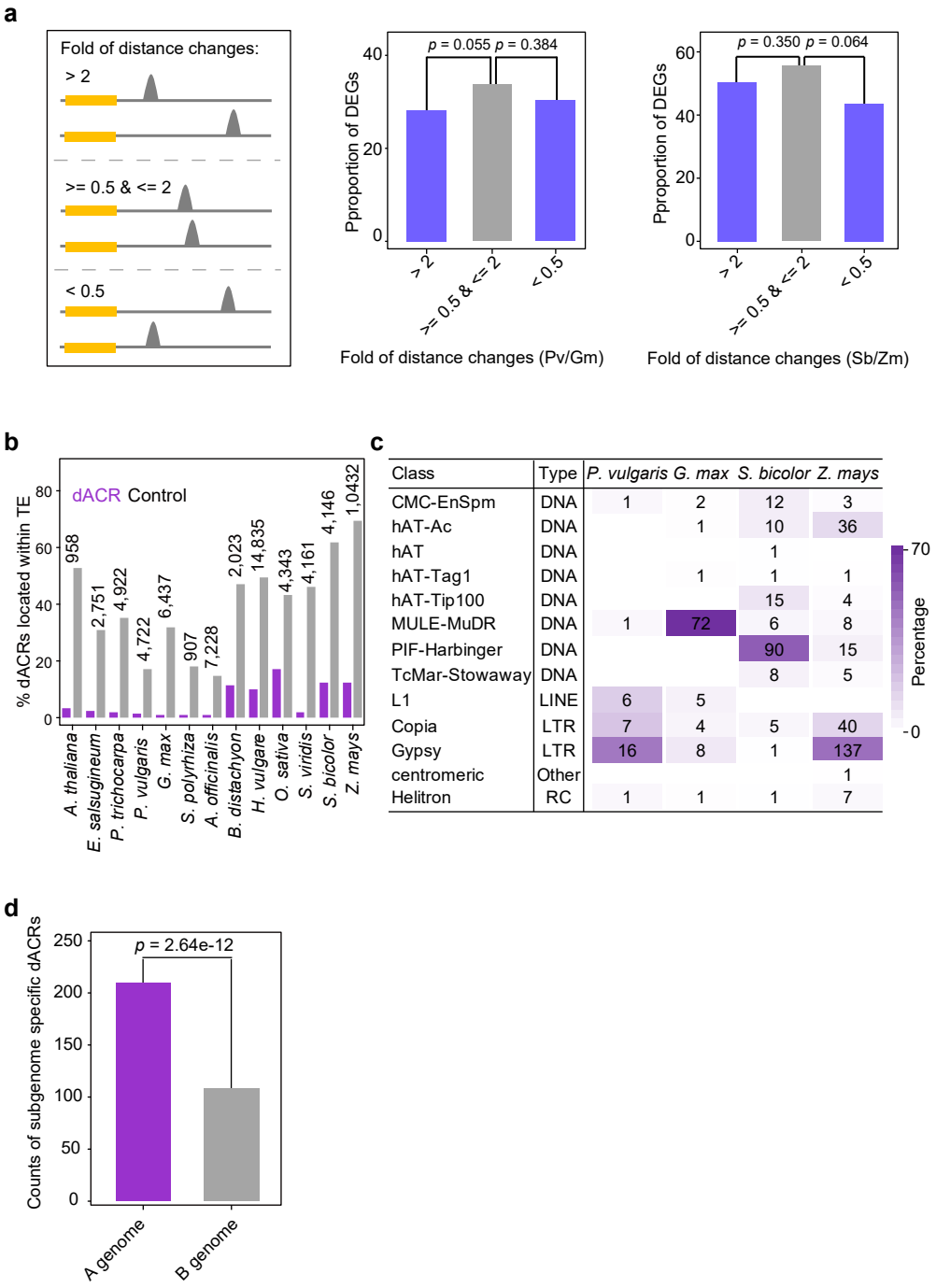
Supplementary Fig. 1: FANS-ATAC-seq to identify plant ACRs. a, Correlation between two biological replicates of ATAC-seq in each species. Each point indicates the relative read density for a given ACR. *R2* values represent percentage of the variation that fits a linear model (*p*-values were calculated by comparing a fitted model to a null model with the *anova* function; *p* - value < 1×10^{-16}). **b,** Toy example demonstrating the enrichment of chromatin accessibilities at known plant enhancers. **c,** Examples of known plant enhancers and their associated genes. **d and e,** Correlation between ACR and counts of all annotated gene (**d**) and expressed gene (**e**) across species. Genes with TPM>0 were considered as expressed genes. At, *Arabidopsis thaliana*; Sp, *Spirodela polyrhiza*; Es, *Eutrema salsugineum*; Bd, *Brachypodium distachyon*; Os, *Oryza sativa*; Sv, *Setaria viridis*; Pv, *Phaseolus vulgaris*; Pt, *Populus trichocarpa*; Sb, *Sorghum bicolor*; Gm, *Glycine max*; Ao, *Asparagus officinalis*; Zm, *Zea mays*; Hv, *Hordeum vulgare*. *R2* values represent percentage of the variation that fits a linear model (n=13; *p*-values were calculated by comparing a fitted model to a null model with the *anova* function; *p* - value < 1×10^{-16}). **f,** Correlation between ACR and gene counts across 500kb non-overlapping windows. *R2* values represent percentage of the variation that fits a linear model (*p*-values were calculated by comparing a fitted model to a null model with the *anova* function; *p* - value < 1×10^{-16}).

Supplementary Fig. 2



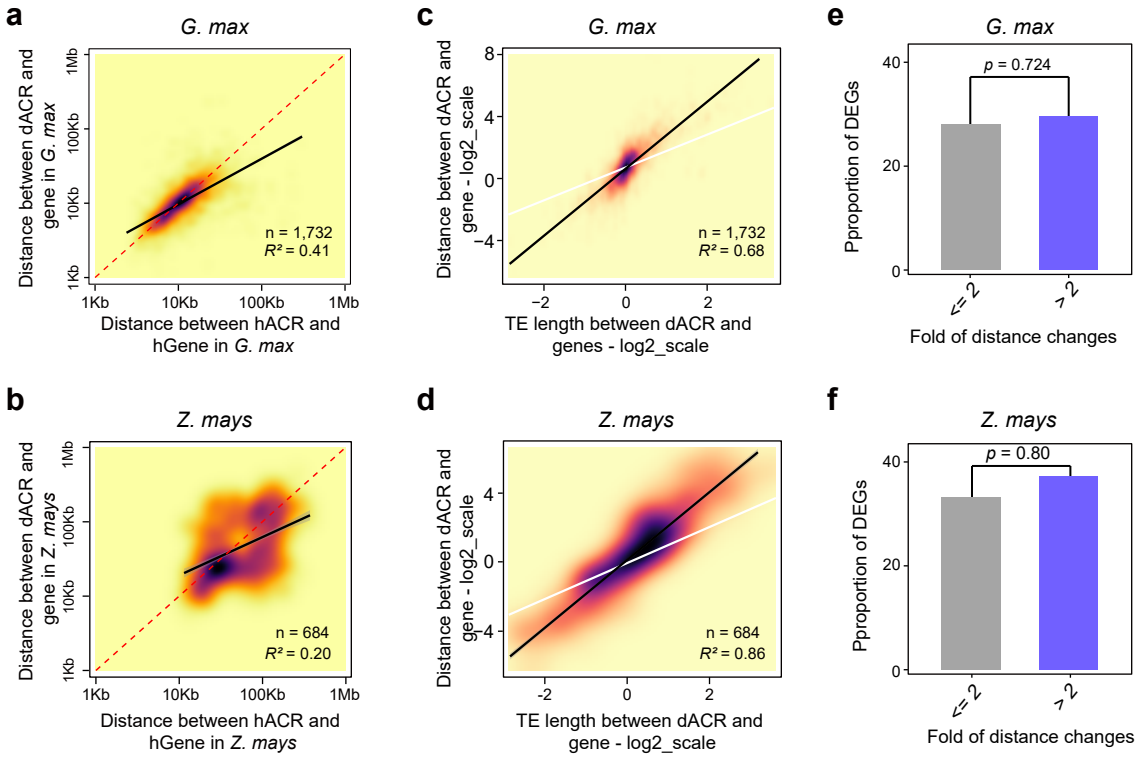
Supplementary Fig.2: Characteristic features of ACRs. a, Metaplots showing the enrichment of ACRs from 5kb upstream to 5kb downstream of transcriptional start sites (TSSs) (top) and transcriptional end sites (TES) (bottom), respectively. **b,** Metaplots showing the enrichment of DNA methylation context (CG, CHG and CHH) from 2kb upstream to 2kb downstream of ACR summits. **c,** GC content of ACRs relative to other intergenic regions across plant species. The box plot displays the distribution of data based on 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 or IQR). The segment inside the rectangle shows the median and "whiskers" above and below the box show the locations of the minimum and maximum. The brown shadows indicate the distribution. The number indicates the samples used in the analysis. K indicates 1×10^3.

Supplementary Fig. 3



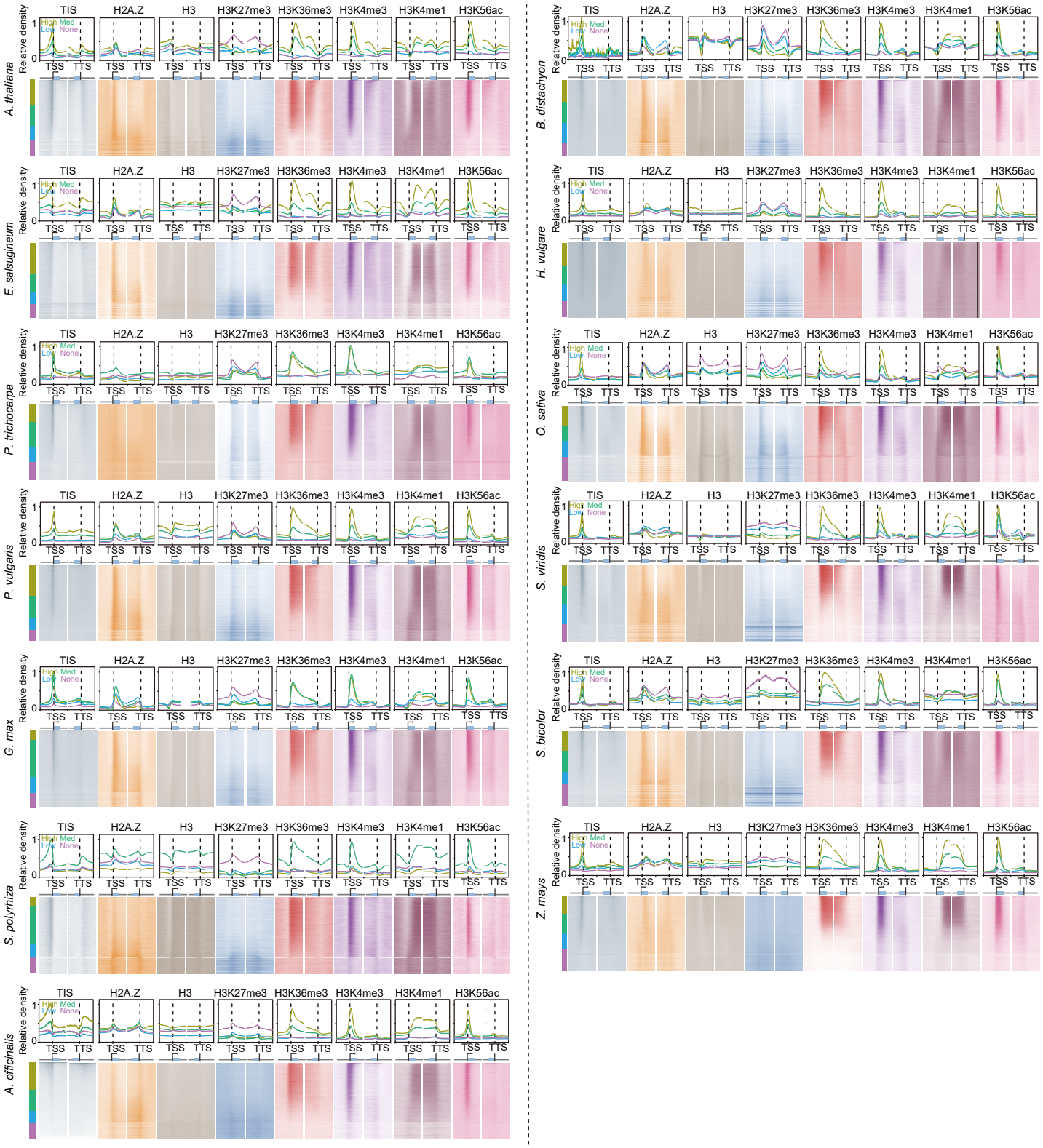
Supplementary Fig. 3: Evolution of dACRs within and between different species. a, Proportions of differentially expressed genes flanked by variable (fold change distance greater than 2 and less than 0.5) and nonvariably (fold change distance between 0.5 and 2) positioned dACRs for comparisons between (left) *P. vulgaris* (Pv)/*G. max* (Gm) (n=1,898) and (right) *S. bicolor* (Sb)/*Z. mays* (Zm) (n=358). Hypothesis testing was conducted with Fisher's exact test (two sided). **b,** Proportion of dACRs embedded within an annotated transposon. The number indicates the sample sizes. All comparisons with control regions were significant (Fisher's exact test; two sided; $p < 0.01$). **c,** The counts of each TE class overlapping with species -specific dACRs. **d,** Prevalence of subgenome-specific dACRs in *Z. mays* subgenome A. *p*-values were calculated by Fisher's exact test (two sided) with a null hypothesis that dACRs are equally likely to be deleted from either chromosome. Total dACRs (n = 10,432) used in the analysis.

Supplementary Fig. 4



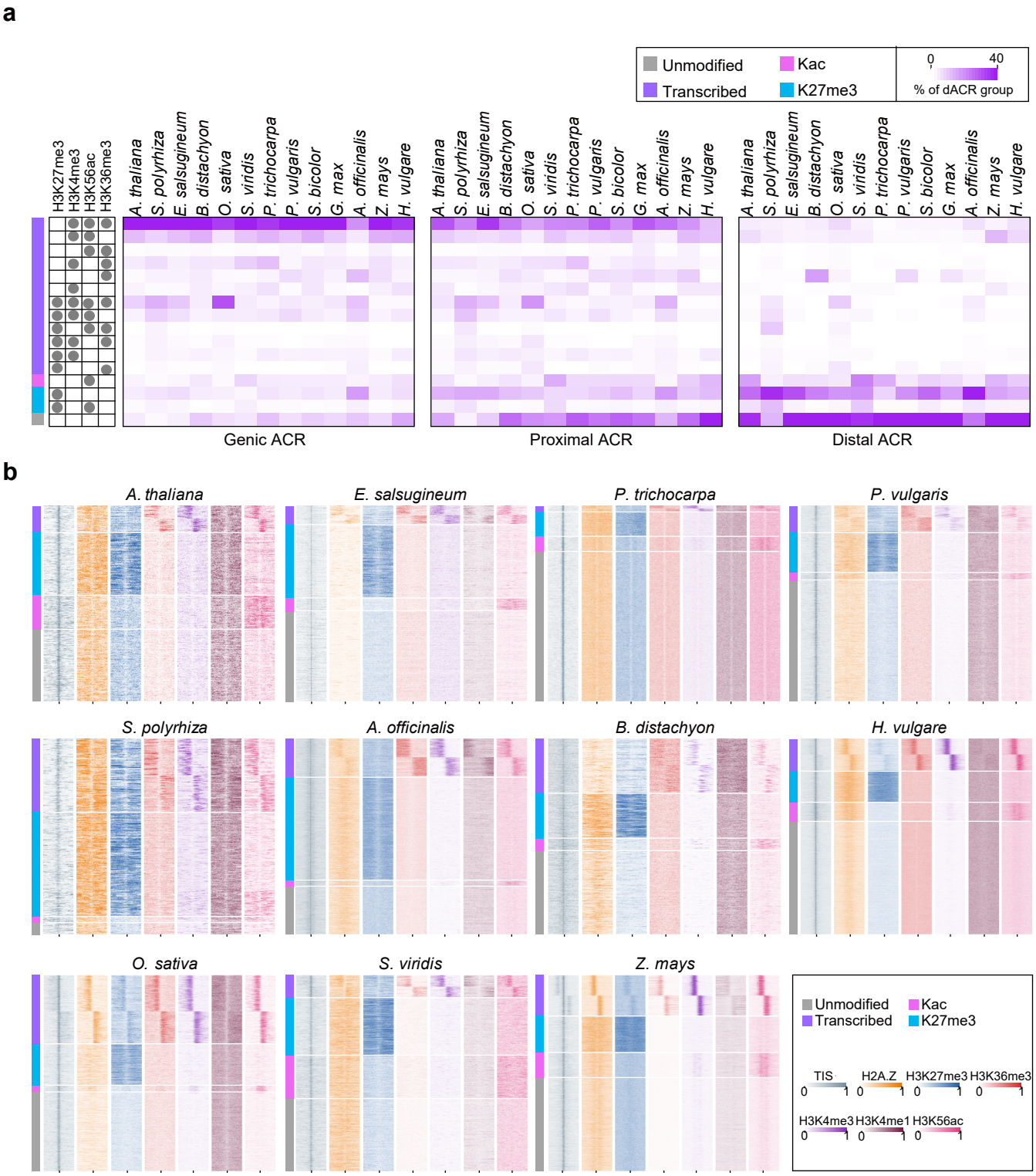
Supplementary Fig. 4: Variation in the distance of dACRs to genes in polyploid events was largely accounted for by TE sequences. a - b, Comparison of dACR and gene intervening distances between duplicated regions for "shared" pairs. *R2* values represent percentage of the variation that fits a linear model (n=13; *p*-values were calculated by comparing a fitted model to a null model with the *anova* function; *p* - value < 1×10^{-16}). The central line indicate the distances (**a**) and ratio (**b**) are equal. Shaded region indicates the standard normal density. **c - d,** Correlation between dACR-gene intervening distances and annotated TE sequence occupancy in each genome. *R2* values represent percentage of the variation that fits a linear model (n=13; *p*-values were calculated by comparing a fitted model to a null model with the *anova* function; *p* - value < 1×10^{-16}). The central line indicates the distances (**c**) and ratio (**d**) are equal. Shaded region indicates the standard normal density. **e - f,** Proportion of DEGs (between paralogs) within *G. max* (**e**; n=1,732) and *Z. mays* (**f**; n=684). Hypothesis testing was conducted using Fisher's exact test (two sided).

Supplementary Fig. 5



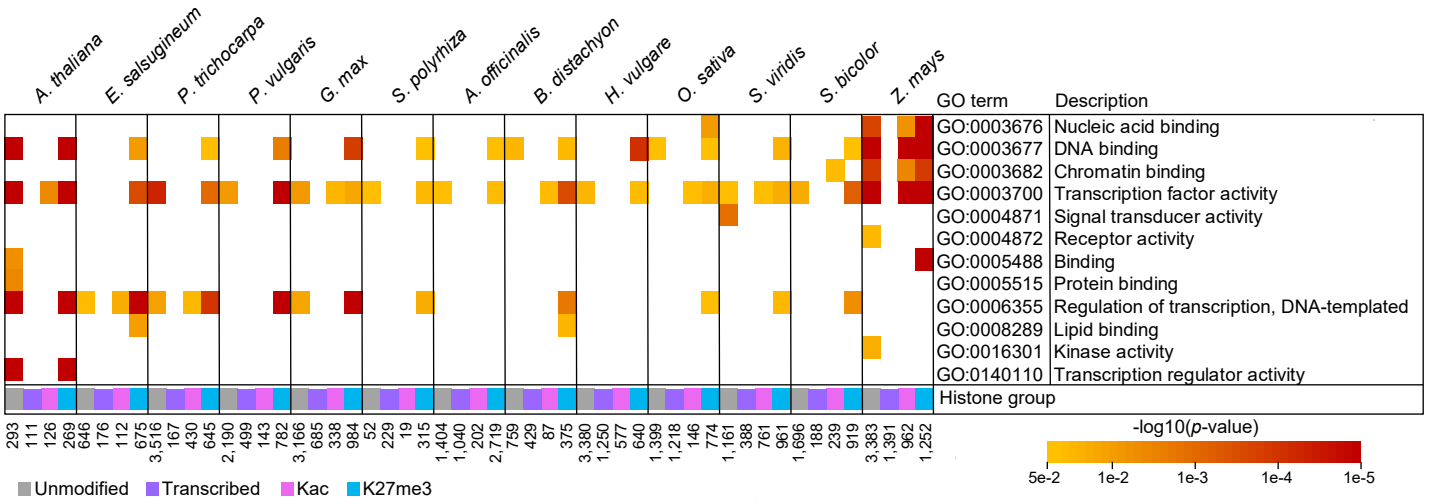
Supplementary Fig. 5: Chromatin accessibility and histone modifications around TSS and TTS in different plant species. Distribution of chromatin modifications and accessibility surrounding genes. Genes are rank ordered by expression (highest to lowest). Metaplots above each heatmap are derived from genes binned by expression level: FPKM greater than 10 (high), FPKM between 1 and 10 (med), FPKM greater than zero and less than 1 (low), and FPKM equal to 0 (none).

Supplementary Fig. 6



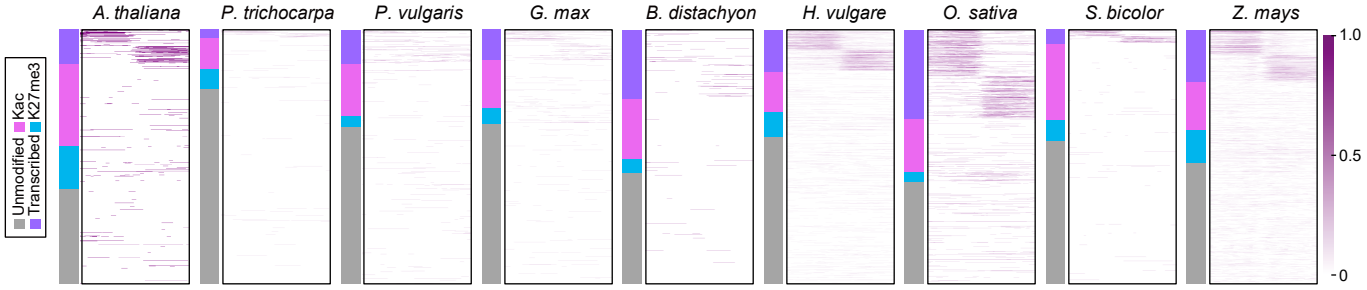
Supplementary Fig. 6: Distinct histone modifications are enriched flanking dACRs. a, Distribution of histone modifications +/- 2kb flanking dACR summits. Clustering was determined by k-means. **b,** Heatmap illustrating the distribution of histone modifications flanking dACRs in different plant species.

Supplementary Fig. 7



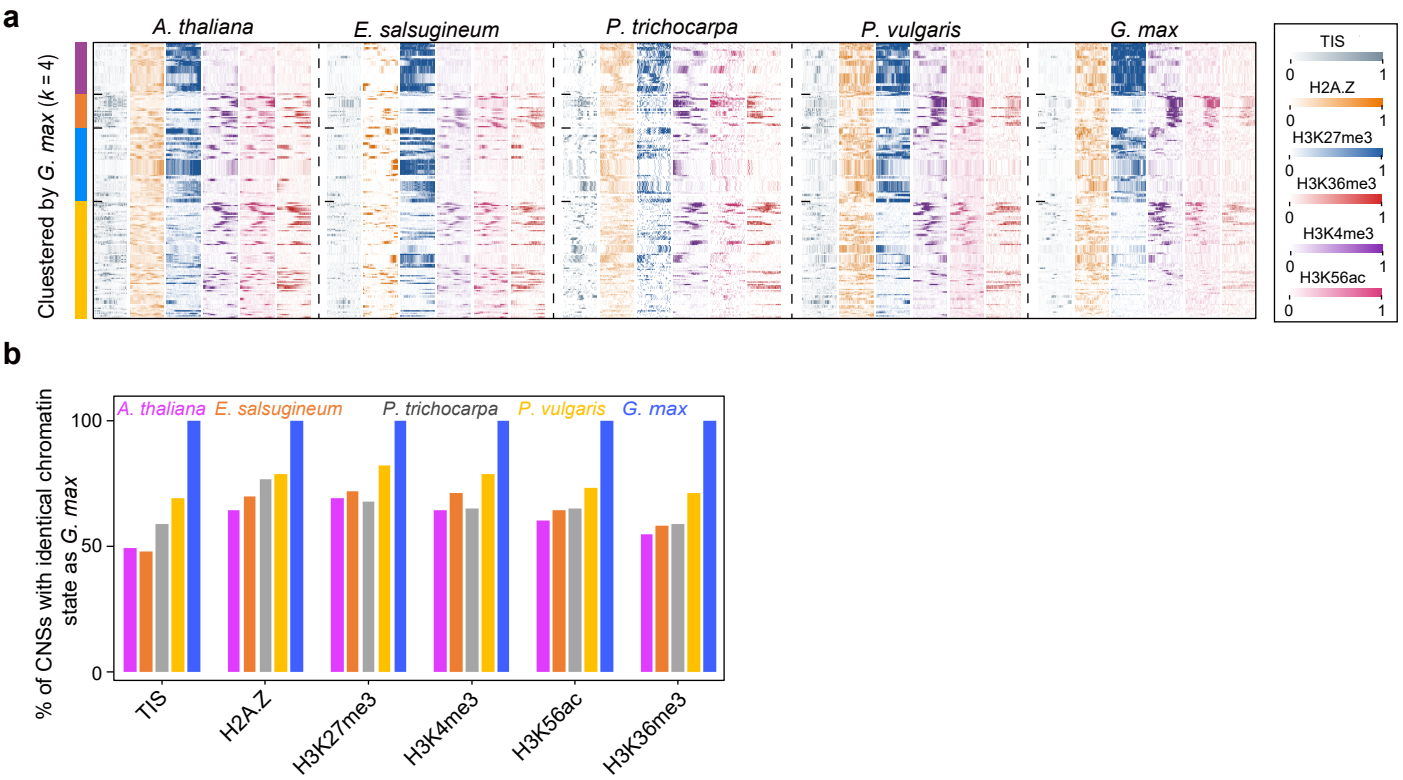
Supplementary Fig. 7: Enriched GO terms of genes near dACRs with distinct histone modifications. Significantly enriched molecular function GO terms for genes flanking different dACR histone classifications across species. Hypothesis testing was conducted with a Hypergeometric Test. *p*-values were adjusted with Benjamini-Hochberg correction. The number in the bottom indicates the gene numbers for the analysis.

Supplementary Fig. 8



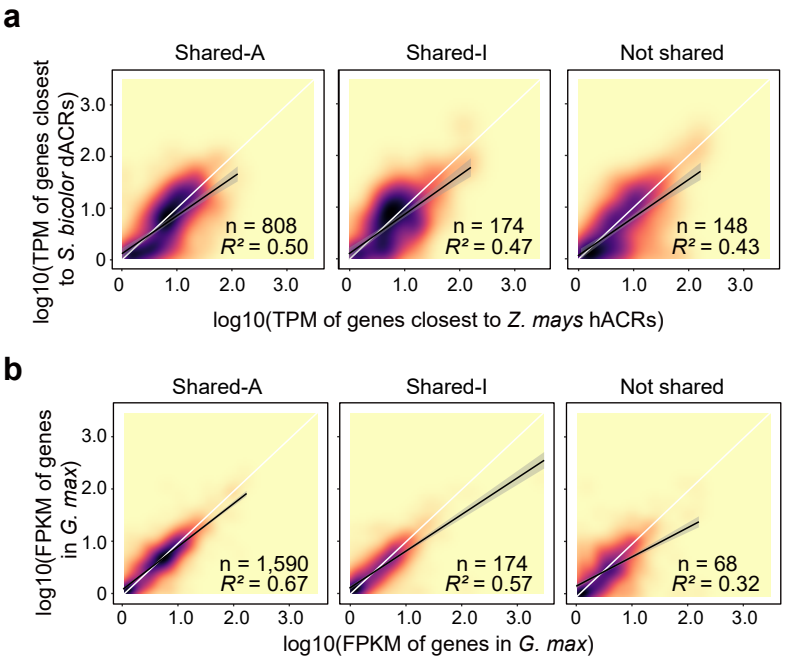
Supplementary Fig. 8: Distribution of ESTs in 4kb windows centered on dACRs from different classifications. The density of EST tags in each 10bp bin was calculated to generate the matrix. *k*-mean ($k = 3$) were used to cluster the matrix.

Supplementary Fig. 9



Supplementary Fig. 9: dACR chromatin states are preserved across eudicots. a, Conservation of chromatin accessibility and histone modifications around CNSs. Colors correspond to relative read densities from ATAC-seq and histone ChIP-seq +/- 2kb of CNS centers clustered using k-means ($k = 4$) in *Glycine max*. Row ordering in the other species were aligned with *Glycine max*. **b,** Proportion of CNSs with an identical chromatin state as *Glycine max*.

Supplementary Fig. 10



Supplementary Fig.10: Divergent expression patterns for genes near species- and subgenome-specific dACRs. Comparison of expression values for the genes closest to a species-specific dACRs between *S. bicolor* and *Z.mays* (**a**) or subgenome-specific dACR within *Glycine max* (**b**). *R2* values represent percentage of the variation that fits a linear model (*p*-values were calculated by comparing a fitted model to a null model with the *anova* function; p - value < 1×10^{-16}). The black line indicates the expression levels are equal and the white line indicates the ratios are equal. Shaded region indicates the standard normal density.