

The regulatory potential of transposable elements in maize

Received: 25 September 2024

Accepted: 11 April 2025

Published online: 13 May 2025

 Check for updates

Kerry L. Bubb^{1,6}, Morgan O. Hamm^{1,6}, Thomas W. Tullius², Joseph K. Min¹, Bryan Ramirez-Corona¹, Nicholas A. Mueth¹, Jane Ranchalis³, Yizi Mao³, Erik J. Bergstrom¹, Mitchell R. Vollger³, Cole Trapnell^{1,4,5}, Josh T. Cuperus^{1,5}, Andrew B. Stergachis^{1,3,4,5}✉ & Christine Queitsch^{1,4,5}✉

The genomes of flowering plants consist largely of transposable elements (TEs), some of which modulate gene regulation and function. However, the repetitive nature of TEs and difficulty of mapping individual TEs by short-read sequencing have hindered our understanding of their regulatory potential. Here we show that long-read chromatin fibre sequencing (Fiber-seq) comprehensively identifies accessible chromatin regions (ACRs) and CpG methylation across the maize genome. We uncover stereotypical ACR patterns at young TEs that degenerate with evolutionary age, resulting in TE enhancers preferentially marked by a novel plant-specific epigenetic feature: simultaneous hyper-CpG methylation and chromatin accessibility. We show that TE ACRs are co-opted as gene promoters and that ACR-containing TEs can facilitate gene amplification. Lastly, we uncover a pervasive epigenetic signature—hypo-5mCpG methylation and diffuse chromatin accessibility—directing TEs to specific loci, including the loci that sparked McClintock's discovery of TEs.

Transposable elements (TEs), first described as ‘controlling elements’ by Barbara McClintock^{1–7}, have the potential to shape the regulation of the host genome^{8–11}. For example, the insertion of a TE in a regulatory region of the maize domestication gene *teosinte branched1* (*tb1*) enhances its expression, contributing to the increased apical dominance of maize compared with its ancestor *teosinte*¹². Although over 80% of the maize genome is annotated as intact TEs or TE fragments¹³, a comprehensive analysis of their regulatory potential is lacking. Commonly used methods to map regulatory elements (that is, accessible chromatin regions, ACRs) have relied on short sequence reads which rarely map uniquely within TEs. Here we use the long-read method Fiber-seq to overcome this limitation and map ACRs across the maize B73 genome. Fiber-seq uses a non-specific DNA *N*⁶-adenine methyltransferase to methylate accessible adenines¹⁴—a modification that is extremely sparse in plants¹⁵, including maize (Extended Data Fig. 1a)—followed by single-molecule PacBio sequencing of ~18 kb maize

chromatin fibres, enabling the synchronous detection of accessible adenines (m6A) and endogenous cytosine methylation (5mCpG).

Assessing the single-molecule regulatory landscape of maize

We compared Fiber-seq and assay for transposase-accessible chromatin using sequencing (ATAC-seq) using paired samples of leaf protoplasts isolated from 14-day-old dark-grown maize seedlings (Fig. 1a and Extended Data Fig. 1b–d). The use of leaf protoplasts minimized cell-type heterogeneity as leaf tissue is enriched in mesophyll cells. We observed that Fiber-seq-derived m6A and 5mCpG calls showed the expected signals at ATAC-seq-derived ACRs and cap analysis of gene expression sequencing (CAGE)-defined transcription start sites (TSSs), in addition to the expected correlation of signal intensity with gene expression at TSSs (Extended Data Fig. 1e–h). However, unlike ATAC-seq, Fiber-seq also revealed periodic m6A signals downstream

¹Department of Genome Sciences, University of Washington, Seattle, WA, USA. ²Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, USA. ³Division of Medical Genetics, Department of Medicine, University of Washington School of Medicine, Seattle, WA, USA. ⁴Molecular and Cellular Biology Program, University of Washington, Seattle, WA, USA. ⁵Brotman Baty Institute for Precision Medicine, University of Washington, Seattle, WA, USA. ⁶These authors contributed equally: Kerry L. Bubb, Morgan O. Hamm. ✉e-mail: absterga@uw.edu; queitsch@uw.edu

of the TSS that were most pronounced for highly expressed genes, reflecting promoter-proximal well-positioned nucleosomes typically measured by micrococcal nuclease sequencing (MNase-seq) (Extended Data Fig. 1f,g)¹⁶.

To rigorously distinguish regions with elevated exogenous m6A signal (methyltransferase-sensitive patches, MSP) due to nucleosome linkers from regions representing ACRs (Extended Data Fig. 1g), we called FIRE elements (Fiber-seq Inferred Regulatory Elements) with the semi-supervised machine learning classifier ‘fiberseq-FIRE’¹⁷. After recalibrating fiberseq-FIRE for maize, 4.6 million methyltransferase-sensitive patches were classified as actuated FIRE elements (precision >0.9), with the remaining 150 million classified as nucleosome linkers. By aggregating single-molecule FIRE elements across the genome, we called 106,867 FIRE ACRs (false discovery rate (FDR) < 0.01, Fig. 1b, and Supplementary Tables 2 and 3). In contrast, we called only 51,817 ACRs with ATAC-seq (*q*-value < 0.01, Supplementary Table 3), consistent with Fiber-seq revealing a more comprehensive regulatory landscape of maize. Fiber-seq identified the vast majority of ACRs called with ATAC-seq in paired samples (Fig. 1b), added ACRs in repeat regions with low mappability, and corrected for false-positive ATAC ACRs, such as those in nuclear genomic regions with homology to plastid or mitochondrial genomes¹⁸ (Fig. 1c,d, and Supplementary Tables 4 and 5). Signal intensity at ACRs in the paired bulk ATAC-seq strongly correlated with the Fiber-seq signal (Extended Data Fig. 1i). However, for a set of ~40,000 shared ACRs, most were detected with Fiber-seq on half or more of the sequenced chromatin fibres, whereas fewer than 5% of cells showed Tn5 insertions in these ACRs in single-cell ATAC-seq¹⁹ (Extended Data Fig. 1j and Supplementary Table 6). This comparison illustrates the limitations of single-cell ATAC-seq as a quantitative measure of per-molecule chromatin accessibility. Taken together, our results show that Fiber-seq accurately captures chromatin accessibility and 5mCpG in maize, with single-molecule and single-nucleotide precision, at a sensitivity twice that of ATAC-seq.

Fig. 1 | Fiber-seq captures the regulatory landscape of maize comprehensively.

a, Experimental scheme. **b**, ACRs called in paired Fiber-seq and ATAC-seq experiments are shown in three bar graphs representing FIRE ACRs (purple) that did not overlap with ATAC ACRs (*n* = 51,878), FIRE ACRs that overlapped with ATAC ACRs (purple/gold, *n* = 54,989), ATAC ACRs that did not overlap with FIRE ACRs (gold, *n* = 3,487), in addition to a bar graph representing shifted control regions (10 kb downstream of FIRE ACRs, *n* = 106,867). ACRs in each category were hierarchically classified as having (1) a FIRE accessibility score of <0.25, (2) a mean FE length <200 bp, (3) low short-read mappability or (4) none of the above (regular ACR). Stacked bar charts indicate the distribution of these classifiers for each ACR category. **c**, Percentage of base pairs (*y* axis) overlapping with ACRs called by either Fiber-seq (FIRE ACRs, triangles) or ATAC-seq (ATAC ACRs, circles) for distinct genomic regions (*x* axis). As expected, fewer base pairs in regions with homology to plastid and mitochondrial genomic sequence overlap with FIRE ACRs (red triangles) than with ATAC ACRs, while more base pairs in regions annotated as transposable elements or repeats overlap with FIRE ACRs (blue triangles) than with ATAC ACRs. **d**, Screenshots of three genomic regions illustrating marked differences between FIRE and ATAC ACR calls. Top to bottom tracks: genomic location, mappability calculated as in Extended Data Fig. 2a, annotated genes, chloroplast sequence, ATAC-seq signal with ATAC ACRs indicated below as rectangles in gold, Fiber-seq signal with FIRE ACRs indicated below as rectangles in purple, FIRE histogram and individual chromatin fibres with FIRE elements in shades of red, with darker shades indicating greater significance. Left: the region highlighted in grey contains two FIRE ACRs but no ATAC ACRs because of low mappability. Middle: the chloroplast sequence track indicates high sequence homology at this nuclear locus with the plastid genome. The ATAC ACR in the highlighted region is a false positive due to incorrect mapping of short sequence reads. Right: the highlighted region shows a FIRE ACR with underlying short FIRE elements. No ATAC ACR was called. ATAC-seq signal is a sliding-window histogram displaying the number of Tn5 insertion sites, with the height of each 20-bp bar representing the number of Tn5 insertions

Short FIRE elements indicate ATAC ACRs in other tissues

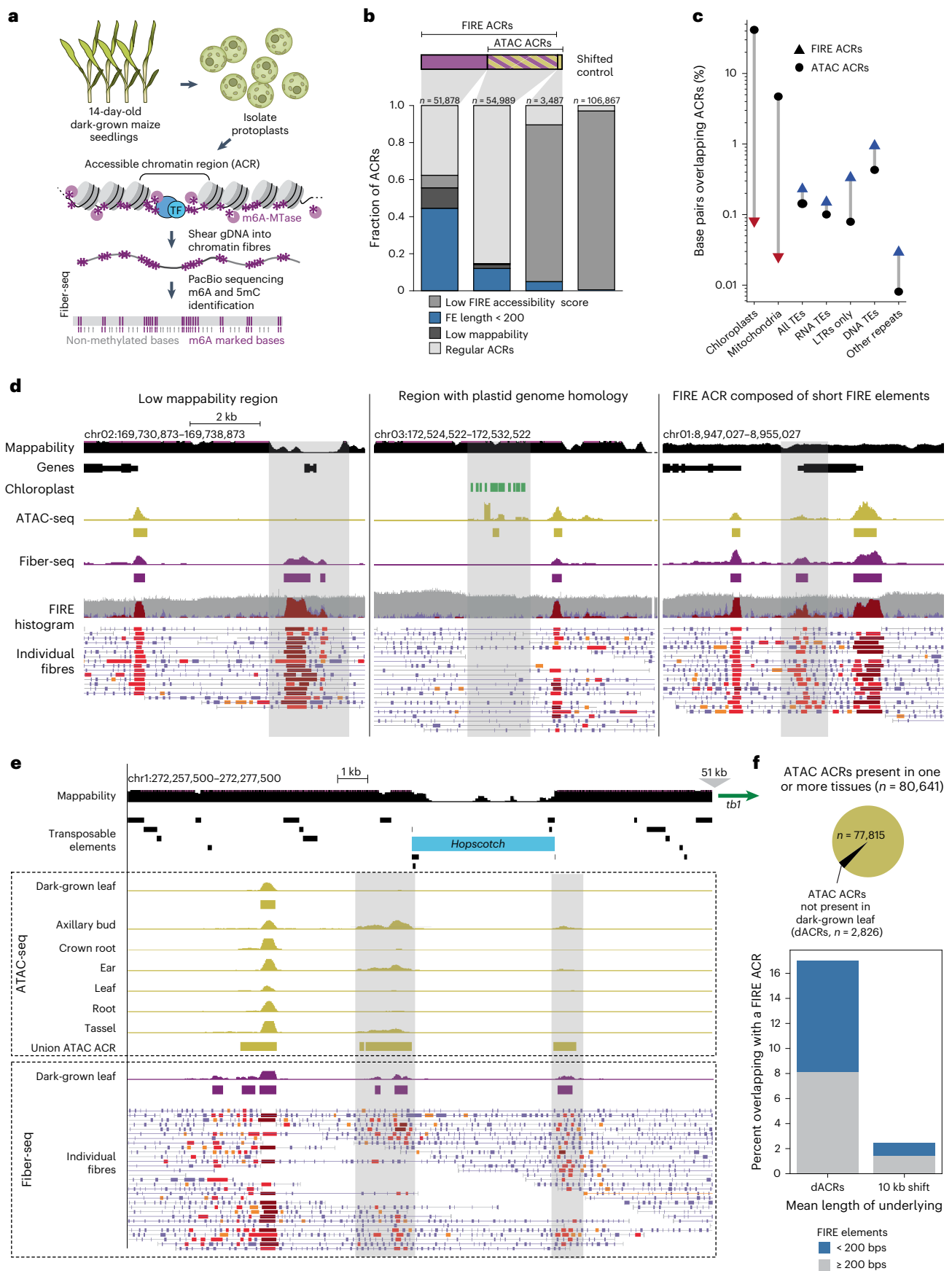
The nearly twice as many ACRs identified by Fiber-seq as compared with the paired ATAC-seq ACRs were only in part explained by low mappability of ATAC-seq reads (Extended Data Fig. 2a,b). Rather, over half of the FIRE ACRs missed by ATAC-seq were composed of short FIRE elements (<200 bp) (Fig. 1b and Extended Data Fig. 2c,d). These FIRE ACRs shared the features typical of ACRs comprising long FIRE elements identified by both methods, such as enrichment of the 6 mA signal, depletion of the 5mCpG signal (Extended Data Fig. 2e) and genomic distribution (Extended Data Fig. 2r).

We detected ACRs composed of short FIRE elements flanking the TE insertion that introduced a *tb1* enhancer¹² but failed to detect these by ATAC-seq (Fig. 1e). However, these flanking regions were detected as ATAC ACRs in embryonic and reproductive tissues (axillary bud, tassel and ear)¹⁹, suggesting that ACRs comprising short FIRE elements may mark genomic loci with tissue-specific chromatin accessibility in maize (Fig. 1e). To systematically evaluate this possibility, we identified ATAC ACRs present in one or more tissues (that is, union ACRs)¹⁹, and then filtered for the subset of these union ACRs for which the corresponding genomic loci showed only background ATAC signal in dark-grown leaves (that is, differential ACRs not present in dark-grown leaves, dACRs), yielding 2,826 dACRs from the total of 80,641 union ACRs (Supplementary Tables 7, 8). Of the 2,826 dACRs, 480 overlapped with a FIRE ACR (17%) and over half of the 480 overlapped with FIRE ACRs comprising short FIRE elements (251/480, Fig. 1f). This result is thus consistent with these ACRs composed of short FIRE elements corresponding to functional regulatory elements in maize that display tissue-selective activity.

Distinctive ACRs mark functional LTR retrotransposons

We next sought to interrogate ACRs in TEs, focusing on long terminal repeat (LTR) retrotransposons because of their prevalence in the maize

within a 150-bp window centred on these 20 bp (minimum = 0, maximum = 210 mapped reads/16,662,983 total mapped reads = 1.26×10^{-5}). ATAC ACRs are MACS2-derived peaks (*q* < 0.01)⁵⁵. Fiber-seq signal is a per-nucleotide average of the scaled, log-transformed ML-model-derived probability of each underlying fibre containing a FIRE element at that nucleotide. Features used by the ML model include m6A density, length of methyltransferase-sensitive patch and A/T content (minimum = 0, maximum = 100). FIRE ACRs are called using ‘fiberseq-FIRE’ (FDR < 0.01)⁵⁰. FIRE histogram is three overlaid histograms: inferred nucleosome *x*-coverage (grey), inferred MSP *x*-coverage (light purple), inferred FIRE element *x*-coverage (red). Individual fibres are annotated with MSPs (light purple) and FIRE elements (reds, FDR ≤ 5%; oranges, 5% < FDR ≤ 10%). **e**, Loci lacking ATAC ACRs in dark-grown maize leaves that show ATAC ACRs in other tissues often overlap with FIRE ACRs composed of short FIRE elements. A screenshot is shown for the region upstream of the *tb1* gene (green arrow, Zm00001eb054440) in which a hopscotch TE insertion (light blue) generated an enhancer. Top to bottom tracks: genomic locus; mappability as in Extended Data Fig. 2a; ATAC-seq data (in first dashed box): ATAC-seq signal (gold) for dark-grown leaves (this study), subsequent tracks pseudobulked single-cell ATAC-seq signal for indicated tissues¹⁹, and union ATAC ACRs (last track) (present in at least one tissue) as golden rectangles; Fiber-seq data (in second dashed box): Fiber-seq signal (purple) in dark-grown leaves and indicated below FIRE ACRs as purple rectangles, individual fibres with short FIRE elements in shades of red. In dark-grown leaves, ACRs were detected by Fiber-seq but not ATAC-seq in the loci flanking the hopscotch TE. However, ATAC ACRs were detected in these loci in axillary bud, ear and tassel tissue. **f**, Of the 80,641 union ATAC ACRs across these seven tissues, 2,826 were not detected in dark-grown leaves (differentially accessible ACRs, dACRs, see Methods for details). About 17% of the loci overlapping with these differentially accessible ACRs overlap with FIRE ACRs, and about half of these are FIRE ACRs composed of short FIRE elements. Statistical analyses and *P* values for Fig. 1b,f are in Supplementary Table 1.



genome (74.4%)¹³. Intact LTR retrotransposons are class I TEs with bilateral LTRs that flank an internal region (Fig. 2a,b). Each of the bilateral LTRs are thought to contain the regulatory elements, promoters and adjacent enhancers that drive expression of the TE genes encoded in the internal region²⁰. LTR retrotransposons mobilize through reverse transcription of their mRNA and integration of the complementary DNA into another genomic location. They are divided into autonomous (which encode the proteins needed for transposition) and non-autonomous (which require proteins encoded by other elements for transposition) LTR retrotransposons. It has been challenging to determine the functional activity of individual LTR retrotransposons because their high sequence identity limits the ability of short-read data to be uniquely mapped to individual LTR retrotransposons^{9,11,20}.

Using Fiber-seq, we mapped ACRs residing within each of the 51,882 intact LTR retrotransposons in the maize genome (Supplementary Table 9), as well as ACRs in solo LTRs (Extended Data Fig. 3 and Supplementary Table 10). Only 2% (941/51,882) of intact LTR retrotransposons contained at least one FIRE ACR entirely within one of their bilateral LTRs (Supplementary Table 9), consistent with widespread epigenetic silencing of maize LTRs by RNA-mediated DNA methylation, a plant-specific pathway that targets TEs²¹. Of the 941 ACR-containing LTR retrotransposons, 21% (201/941) contained two adjacent ACRs (paired ACRs, Fig. 2a and Supplementary Table 9) in one or both of their LTRs. Furthermore, 94 of these contained the two adjacent ACRs in both LTRs (paired bilateral ACRs, Supplementary Table 9).

The paired bilateral ACRs almost always exhibited single-molecule co-accessibility and hypo-5mCpG methylation (Fig. 2a). Given their position relative to the 5' end of the transposon and the putative transcription start site, these ACRs probably correspond to the putative LTR promoter and enhancer elements²⁰. Consistent with this assumption, the putative promoters showed far higher predicted promoter scores than the putative enhancers when using a validated convolutional neural net model for promoter strength²² (Extended Data Fig. 4a). Furthermore, the putative LTR promoter and enhancer elements are enriched in distinct transcription factor (TF) motifs, consistent with them having distinct regulatory roles (Supplementary Tables 11–14).

To further validate whether the putative LTR enhancers and promoters have distinct regulatory roles, we applied FiberHMM²³ to call protein occupancy footprints within them. Previous studies using Fiber-seq show that eukaryotic RNA polymerases display MTase footprints of 40–60 bp size, which is distinct from those of TFs, which are typically <40 bp in size^{11,24}. Using this approach, we observed that LTR putative promoters were uniquely marked by MTase footprints of 40–60 bp size (Fig. 2c), consistent with these elements being RNA polymerase-bound promoters. In contrast, the LTR putative enhancers were largely marked by MTase footprints of <40 bp size that were well aligned with enriched TF motifs (Fig. 2c,d), consistent with them being occupied by TFs. The putative enhancer and promoter ACRs were generally separated by 3 or 4 well-positioned mononucleosome footprints, with the regions upstream and downstream of the putative promoter showing larger, less clearly phased di- and trinucleosome footprints, typically associated with heterochromatin (Extended Data Fig. 5a). Although the putative enhancer and promoter ACRs were nearly constitutively accessible on all fibres, protein occupancy by 40–60 bp footprints and <40 bp footprints varied substantially between them (Fig. 2c).

Together, these findings demonstrate that Fiber-seq enables the identification of individual LTRs within the maize genome that contain active chromatin at both the putative enhancers and promoters, consistent with these LTRs being poised to be functionally active. In total, in maize leaves there are only 94 LTR retrotransposons that contain the functional regulatory elements required for transposon mobilization, with only 76 of these being autonomous LTR retrotransposons.

Single LTR ACRs are putative enhancers

Most LTR retrotransposons that harbour an ACR contained only a single ACR in one or both of their LTRs (Fig. 2b and Supplementary Table 9). Intact LTR retrotransposons with single ACRs were enriched for containing a single ACR in both of their bilateral LTRs (that is, single bilateral ACRs, 499/941, 53%). The single ACRs exhibited far greater single-molecule heterogeneity than paired ACRs (Extended Data Fig. 4b). Specifically, while paired ACRs showed a bimodal actuation distribution with over half being supported by FIRE elements called in 75%

Fig. 2 | FIRE ACRs in intact LTR retrotransposons identify functional retrotransposons.

a, Representative example of an intact LTR retrotransposon with paired FIRE ACRs both in the left and the right long terminal repeats (paired bilateral ACRs in LTRs, ID: LTRRT_14411). Top indicates left and right LTR in blue and paired bilateral FIRE ACRs in red. First and second paired ACRs are labelled, representing the putative enhancer and the putative promoter, respectively. Putative transcription start site is indicated with black arrow, and genes in internal region are indicated. Top to bottom tracks: genomic location, mappability calculated as in Extended Data Fig. 2a, ATAC-seq signal in gold, Fiber-seq signal in purple with FIRE ACRs indicated as purple rectangles below, 5mCpG methylation as the per-CpG methylation probability calculated by pb-CpG-tools⁶⁴. The vertical axis of the 5mCpG track represents the methylation probability at individual CpG sites, expressed as a percentage, minimum = 0, maximum = 100. The CpGs/kb track represents a sliding-window histogram displaying the number of CpG dinucleotides, with the height of each 100-bp bar representing the number of CpG dinucleotides within a 1-kb window centred on these 100 bp. Paired bilateral ACRs in LTR retrotransposons tended to be hypomethylated as expected for accessible regions. **b**, Representative example of an intact LTR retrotransposon with one FIRE ACR both in the left and the right LTR (single bilateral ACRs in LTRs, ID: LTRRT_8308). Tracks as in **a**. The single FIRE ACRs in this LTR retrotransposon showed high levels of 5mCpG methylation coinciding with the m6A signal; magnified detail below shows methylated 5mCpGs (red) and unmethylated CpGs (blue) and m6A methyltransferase-sensitive patches (purple) on individual fibres. **c**, Plot depicting the aggregate percent occupancy of nucleosome-sized footprints (>90 bp, green), putative polymerase-sized footprints (40–80 bp, dark grey) and putative TF-sized footprints (10–40 bp, azure blue) within the putative promoters of all paired ACRs in LTRs. On the x axis, zero indicates a selected base proximal to the 3' end of the putative promoter (typically -75 bp upstream) that is highly conserved

across all putative promoters. **d**, Representative example of footprinted Fiber-seq reads at a paired ACR (zero: chr05:195,444,369, reverse strand). Top: tracks showing the count of subnucleosomal sized footprints (<80 bp, dark blue) and nucleosome-sized footprints (>90 bp, green) at each position normalized to the maximum count within the region. Middle: blocks showing the genomic position of enriched TF motifs, coloured by identity. Bottom: individual Fiber-seq reads with footprints represented by blocks coloured by predicted identity (accessible, thin black line; nucleosome, green; unknown, grey; TF, coloured corresponding to overlapped motif with multiple overlapped motifs indicated via stripes). Reads are clustered and sorted via *k*-means clustering with 3 clusters, indicated via a column of colours at right. **e**, LTR retrotransposons with single bilateral FIRE ACRs tended to maintain the ACRs marking putative enhancers. Histogram of FIRE ACR location relative to the 5' edge of a given retrotransposon, stratified by type of ACR. **f**, Plot depicting the aggregate percent occupancy of nucleosome-sized footprints (>90 bp, green), putative polymerase-sized footprints (40–80 bp, dark grey) and putative TF-sized footprints (10–40 bp, azure blue) at positions surrounding the centre of all hypo-5mCpG-methylated single ACRs. **g**, Representative example of footprinted Fiber-seq reads at a single hypomethylated ACR (chr04: 170,173,642–170,173,927, forward strand), with position indicated relative to the centre of the ACR. Tracks as in **d**. **h**, Plot depicting the aggregate percent occupancy of nucleosome-sized footprints (>90 bp, green), putative polymerase-sized footprints (40–80 bp, dark grey) and putative TF-sized footprints (10–40 bp, azure blue) at positions surrounding the centre of all hyper-5mCpG-methylated single ACRs. **i**, Representative example of footprinted Fiber-seq reads at a single hypermethylated ACR (chr04: 206,713,520–206,713,627, forward strand), with position indicated relative to the centre of the ACR. Tracks as in **d**, with the motif track omitted due to a lack of motifs. **j**, Single FIRE ACRs in LTR retrotransposons were more likely to be 5mCpG methylated than paired bilateral FIRE ACRs, regardless of their position.



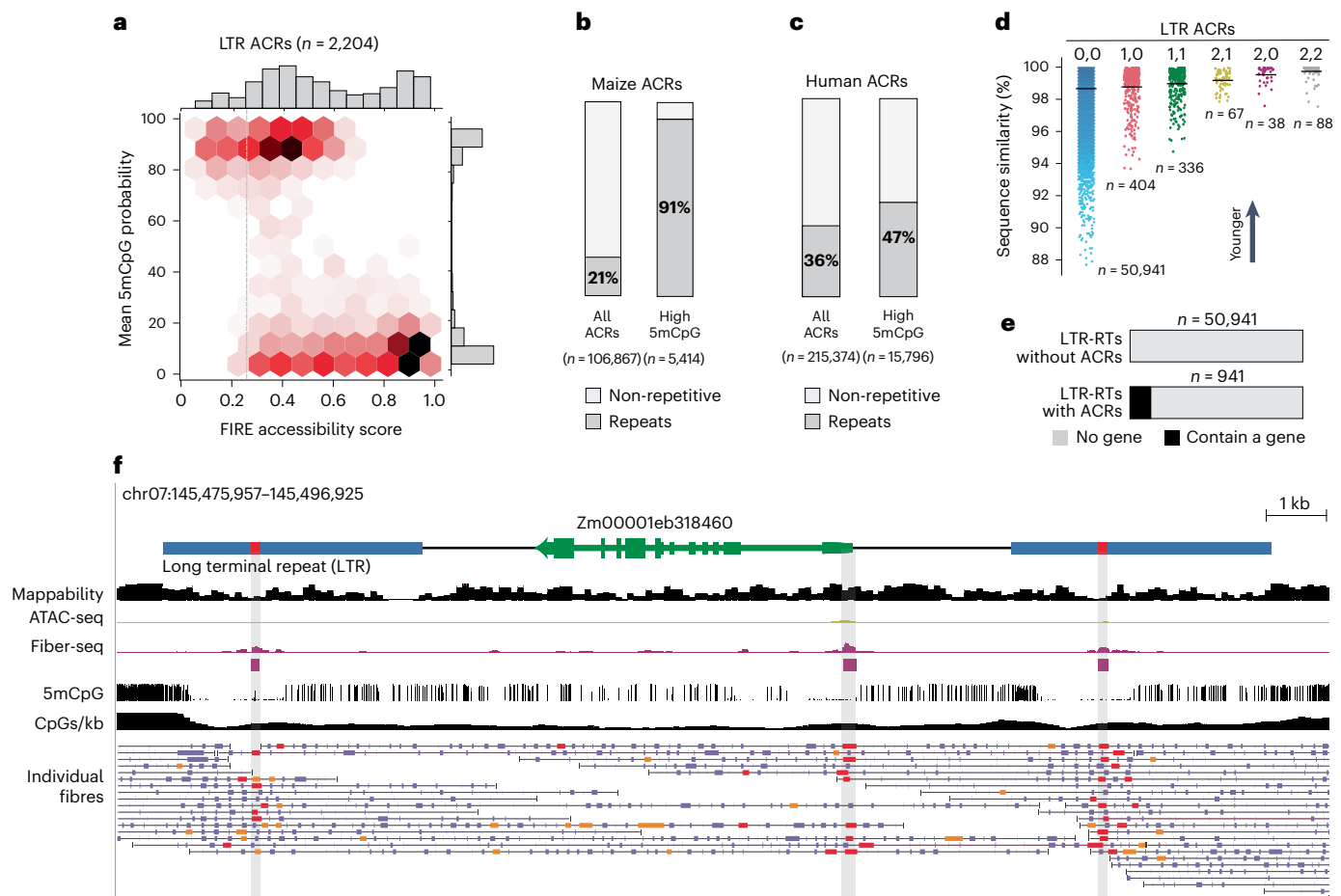


Fig. 3 | Features of FIRE ACRs in repetitive elements. **a**, Hexbin plot shows FIRE accessibility scores (x axis) and mean 5mCpG probabilities (y axis) for 2,204 ACRs in LTRs. Shades of red denote frequency, also shown in plotted histograms (top and right). Of FIRE ACRs in LTRs, 57% (1,261/2,204) were highly 5mCpG methylated and 37% (733/1,978) of high-confidence FIRE ACRs within LTRs (FIRE accessibility score >0.25, grey dashed line) were highly 5mCpG methylated. **b**, Percentage of all FIRE ACRs and FIRE ACRs with high 5mCpG methylation (mean 5mCpG methylation >50%) that overlap with an annotated repeat by >50 bp. **c**, Fraction of all human FIRE ACRs and human FIRE ACRs with high 5mCpG methylation (mean CpG methylation of >50%) that overlap an annotated repeat by >50 bp. FIRE ACR calls from human cell line GM12878 (ref. 43). **d**, The presence of FIRE ACRs correlates with the sequence similarity of left and right

LTRs, a measure of evolutionary age. LTRs with paired bilateral FIRE ACRs showed the greatest sequence similarity, while those without FIRE ACRs showed the least. 0,0, no ACRs; 1,0, single unilateral ACR; 1,1, single bilateral ACRs; 2,1, paired ACR in one LTR, single in the other; 2,0, paired ACR in one LTR, none in the other; 2,2, paired bilateral ACRs. Rare instances of other configurations are omitted. **e**, Fraction of intact LTR retrotransposons with or without at least one FIRE ACR that contain an annotated gene. **f**, The highly duplicated, well-annotated gene (Zm00001eb318460, green) within an LTR retrotransposon is a candidate for TE-enabled gene amplification. Tracks as in Fig. 2a. There are single bilateral ACRs present in the LTRs, in addition to a FIRE ACR marking the transcription start site of this gene (highlighted in grey). Statistical analyses and *P* values for Fig. 3b–e are in Supplementary Table 1.

of underlying fibres, only 7% of single ACRs crossed this actuation threshold (Extended Data Fig. 4b).

We next examined whether single ACRs preferentially localized to the putative LTR enhancer or to the putative LTR promoter. To accomplish this, we first examined the distance of the ACR to the 5' LTR edge, which could be measured within the subset of 268 autonomous LTR retrotransposons containing an ACR, as strandedness could be inferred at these sites. We observed that nearly all LTR retrotransposons with a single ACR selectively retained the ACR that positionally corresponds to the putative LTR enhancer element (Fig. 2e), suggesting that chromatin accessibility is lost at the position of the putative LTR promoters.

Single ACRs contained predicted TF motifs that were more similar to those enriched in the putative LTR enhancers than to those enriched in the putative LTR promoters of paired ACRs (Extended Data Fig. 4c and Supplementary Tables 11–14). Consistently, the single ACRs showed small TF footprints (<40 bp) that were well aligned with the enriched TF motifs (Fig. 2f,g and Supplementary Tables 11–14). Per-fibre chromatin accessibility at these single ACRs was more heterogeneous than that observed at putative enhancers of paired ACRs (Fig. 2f–i and Extended

Data Fig. 4b). However, the occupancy of small footprints along accessible fibres within these single ACRs was consistent with that of the putative LTR enhancers of paired ACRs (Fig. 2f–i).

In stark contrast to paired ACRs in LTRs, or ACRs elsewhere in the maize genome, we observed that a subset of single ACRs in LTRs exhibited hyper-5mCpG methylation coinciding with chromatin accessibility (Figs. 2b,j and 3a), two epigenetic marks thought to be mutually exclusive. Leveraging the single-molecule nature of our chromatin accessibility and 5mCpG methylation calls, we demonstrated that chromatin accessibility and hyper-5mCpG methylation co-occurred and overlapped along the same chromatin fibre at these single ACRs of LTR retrotransposons (Fig. 2b and Supplementary Table 15). Hyper-5mCpG-methylated ACRs and hypo-5mCpG-methylated ACRs showed similar TF footprint occupancy in accessible fibres (Extended Data Fig. 5b).

As expected, hyper-5mCpG methylation was rarely seen overlapping FIRE ACRs elsewhere in the maize genome. The rare ACRs with simultaneous hyper-5mCpG methylation and chromatin accessibility were almost exclusively present within repeat elements (Fig. 3b), with

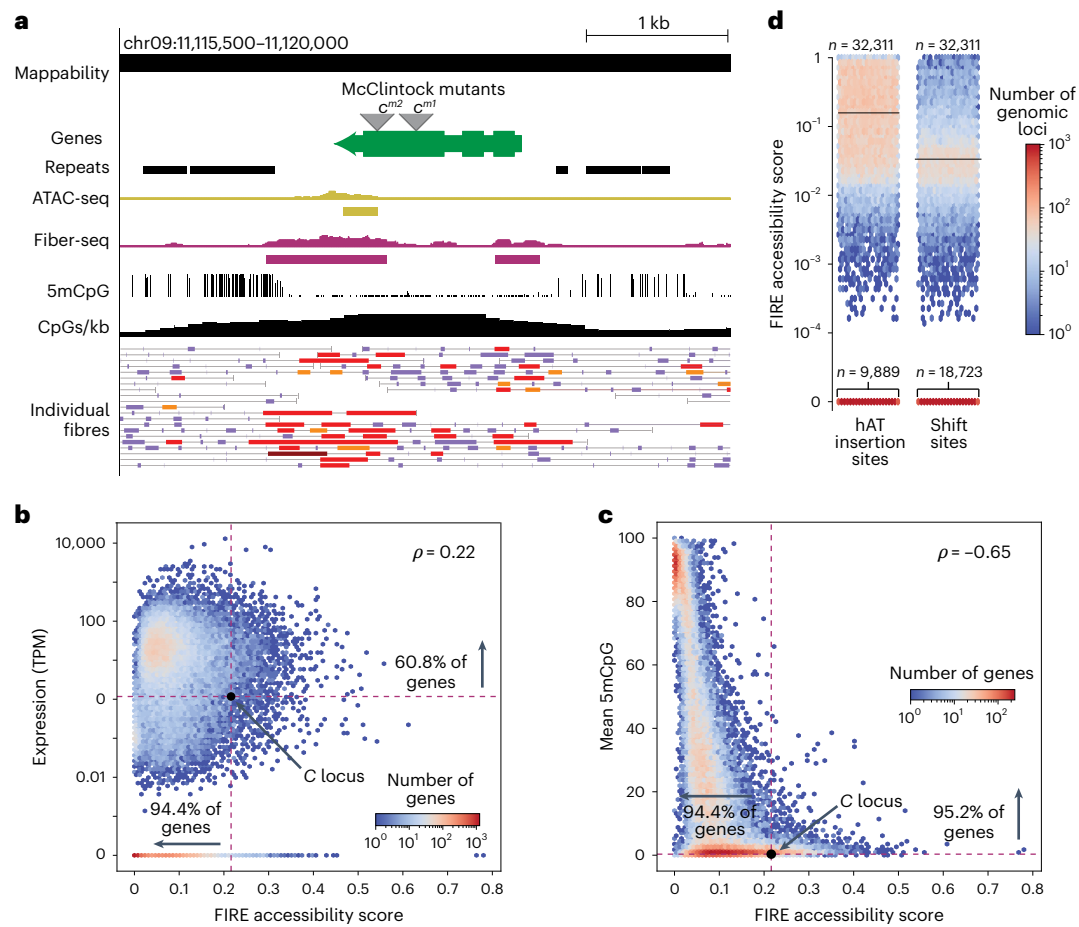


Fig. 4 | hAT TEs tend to insert in regions with diffuse chromatin accessibility detected by Fiber-seq. a, Screenshot of the locus containing the *CI* gene (green, Zm00001eb373660). Tracks as in Fig. 2. hAT TE insertions identified by McClintock as mutant alleles c^{m1} and c^{m2} . The *CI* gene in B73, which does not contain a hAT TE, showed diffuse gene body accessibility detected in Fiber-seq (purple) coinciding with unusual gene body hypo-5mCpG methylation. **b**, The *CI* gene body showed higher FIRE accessibility scores than 94.4% of other genes while showing only slightly above average gene expression. Dashed red lines signify location of C locus. **c**, The *CI* gene body showed lower 5mCpG gene body

methylation than 95.2% of genes. **d**, Loci identified as hAT TE insertion sites in exactly 1 of the 25 non-B73 NAM strains⁴⁵ were more likely to show diffuse gene body accessibility (shown as FIRE accessibility scores) than control loci (same-size regions 10 kb shifted to the right in the B73 genome). Mean and median FIRE accessibility scores for hAT TE insertion sites were 0.159 and 0.046. Mean and median FIRE accessibility scores for shifted control regions were 0.035 and 0. Horizontal lines indicate mean FIRE accessibility scores. Thousands of hAT TE insertion sites showed FIRE accessibility scores of 0 ($n = 9,889$). Statistical analysis and P value for Fig. 4d are in Supplementary Table 1.

15% of these corresponding to single ACRs in intact LTR retrotransposons. In humans, this unexpected co-occurrence of chromatin accessibility and hyper-5mCpG methylation was also rare but not substantially enriched in repeats (Fig. 3c). These results point to the plant-specific RNA-mediated DNA methylation pathway as contributing to this unusual co-occurrence of these two epigenetic marks in maize. However, further analysis of the methylation signatures typical of RNA-mediated DNA methylation or other chromatin states at these low-mappability loci was not feasible because the publicly available data sets resulted from short-read sequencing²⁵.

Given the distinct chromatin features of LTRs containing paired and single ACRs, we reasoned that LTR retrotransposons containing the former might be evolutionarily younger TEs, while TEs containing the latter might be older but still younger than the many fully silent transposons. To address evolutionary age, we examined the sequence similarity between the left and the right LTR of each intact LTR retrotransposon as a metric reflecting time since transposition. LTR retrotransposons with exactly one FIRE ACR per LTR (single ACR) showed greater mean sequence similarity than those without FIRE ACRs (99.0% versus 98.7%, $P = 0.004$, Mann–Whitney U -test) (Fig. 3d). LTR retrotransposons with exactly two FIRE ACRs per LTR (paired ACRs) showed a mean sequence similarity of 99.8%, significantly greater

than those with one FIRE ACR per LTR ($P = 6.8 \times 10^{-26}$, Mann–Whitney U -test) (Fig. 3d). Thus, recently transposed LTR retrotransposons have a characteristic chromatin accessibility and 5mCpG methylation pattern that degenerates with evolutionary age.

LTRs can contain active promoters and propagate host genes

TEs have been long thought to shape host gene regulation by adding or disrupting promoters, enhancers, insulators and coding regions^{25,26}. In humans, TE-derived promoters have been inferred via mapping of transcription start sites²⁷ and TF binding sites overlapping TE sequences^{28,29}. However, in maize, attempts to infer the regulatory effects of LTR retrotransposons have largely been limited to studying gene expression patterns associated with the presence or absence of neighbouring polymorphic TEs that overlap ATAC ACRs²⁸—an analysis that is severely limited by short-read mappability issues inherent to TEs. We sought to leverage our comprehensive maps of FIRE ACRs across intact maize LTR retrotransposons to identify LTRs that may be shaping host gene regulation (Extended Data Fig. 6a,b). Using this approach, we discovered that the putative target gene impacted by LTR FIRE ACRs often resided within the LTR retrotransposon itself. In fact, of the 941 LTR retrotransposons with ACRs in B73, 114 (12%) contained an annotated gene within

the intact retrotransposon (Extended Data Fig. 6c,d), 24-fold greater than LTR retrotransposons without ACRs (Fig. 3e). Of these 114, the promoters of 49 annotated genes were marked by a FIRE ACR, with 48 co-opting one of the LTR ACRs as their promoter (Extended Data Fig. 6d). Overall, these findings indicate that a major way that LTRs shape maize host gene regulation is by providing a novel LTR-encoded promoter element.

In one case, the internal gene (Zm00001eb318460) maintained its promoter, marked by a FIRE ACR, in addition to single FIRE ACRs in the flanking bilateral LTRs (Fig. 3f). This histone deacetylase complex gene is highly expressed (79% percentile in dark-grown maize leaves)³⁰ and has orthologues in the close maize relative *Sorghum bicolor*, its ancestor teosinte (*Zea mays* ssp. *mexicana*)³¹ and other grasses. In human, Alu TEs have been proposed to enable segmental duplication³², so we sought to evaluate whether the gene's residence within the LTR retrotransposon might be associated with its duplication within the B73 genome. Consistent with this hypothesis, we found numerous Zm00001eb318460 paralogues in the B73 genome (21 amino acid blast hits with e -value $< 1 \times 10^{-50}$). This is a highly unusual level of gene duplication for maize genes with a similar expression level and length, as 93% of similar genes showed < 5 amino acid blast hits (Supplementary Table 16). Taken together, these findings provide evidence that TEs enable gene amplification in maize.

Diffuse chromatin accessibility marks insertions of hAT TEs

In general, transposons have been shown to preferentially insert into accessible chromatin both in vitro and in vivo^{33,34}. However, the epigenetic features that predispose genomic loci to insertion of class II (DNA) TEs, in particular hAT TEs, remain unresolved. The first gene reported by McClintock to be susceptible to insertion of hAT TEs is the *C locus*^{33–37}, now called *C1* or *coloured aleurone 1* gene. The *C1* gene body showed unusual diffuse Fiber-seq chromatin accessibility (Fig. 4a), ranking among the top 5% of all genes (Fig. 4b). We also observed hypo-5mCpG methylation across the *C1* gene body, ranking among the bottom 5% of all genes (Fig. 4c). Other genes identified by McClintock as having hAT TE insertions also showed diffuse Fiber-seq chromatin accessibility and hypo-5mCpG methylation within their respective gene bodies (Extended Data Fig. 7)^{35–39}. While diffuse gene body chromatin accessibility was weakly correlated with gene expression (Fig. 4b), this feature was more strongly anticorrelated with hypo-5mCpG methylation (Fig. 4c)^{40–44}. These findings suggest that both diffuse gene body chromatin accessibility and hypo-5mCpG methylation may uniquely mark preferred hAT TE insertion sites. To test this hypothesis, we identified over 32,000 loci in the B73 maize reference genome that contain hAT TE insertions in exactly 1 of the 25 non-B73 NAM founder lines⁴⁵. Indeed, the hAT TE insertion sites were substantially more accessible in B73 than control regions (Fig. 4d and Supplementary Table 17) and were preferentially marked by hypo-5mCpG methylation (Supplementary Table 17). We thus show that the pervasive epigenetic marks of diffuse Fiber-seq chromatin accessibility and hypo-5mCpG methylation guide the insertional landscape of hAT TEs in maize, including those initially described by McClintock.

Discussion

The contemporary maize genome is largely composed of repetitive DNA, with the crop's ~40,000 genes clustered in small islands of non-repetitive sequence⁴⁶. The promise of long-read sequencing for probing regulatory activity in repetitive plant genomes has been illustrated by previous studies in maize⁴⁶ and *Arabidopsis*⁴⁷. However, the reduced accuracy associated with these previous studies' reliance on single-pass long-read sequencing⁴⁸ has limited their ability to perform an in-depth de novo characterization of individual regulatory elements and their associated single-molecule protein occupancy patterns within transposons. Using Fiber-seq paired with PacBio HiFi

sequencing in dark-grown maize leaves, we perform de novo identification of ACRs in maize, identifying twice as many ACRs as compared with using ATAC-seq in paired samples. This dramatic increase in ACR calls is primarily due to the experimental issues with ATAC-seq for detecting short regulatory elements. The ACRs specific to Fiber-seq share the genomic and epigenetic features of ACRs identified by both methods. Furthermore, they often mark loci that show a strong ATAC-seq signal in other tissues. We attribute this phenomenon of 'ACR foreshadowing' to the less frequent and more stochastic TF occupancy at these sites in leaf tissue, which can be captured by Fiber-seq due to its single-molecule resolution. By contrast, the more frequent and more consistent TF occupancy of these same sites in a different tissue results in ACRs that are detectable by both methods.

Unencumbered by mappability concerns, we discover that only 2% (941/51,882) of intact LTR retrotransposons in the maize genome show evidence of regulatory activity, highlighting the efficiency of the plant-specific RNA-mediated DNA methylation pathway to silence the vast majority of transposons in maize. Furthermore, only 94 intact LTR retrotransposons in the maize genome show chromatin accessibility in LTRs in their putative enhancer and promoter elements, chromatin features presumably essential for enabling transposon mobilization.

The presence and specific configurations of ACRs in LTR retrotransposons is correlated with the evolutionary age of that LTR retrotransposon, with accessibility increasingly lost in older transposons. LTR retrotransposons that lose chromatin accessibility preferentially maintain an accessible putative enhancer element. However, the epigenetic pattern at these putative LTR enhancer elements markedly diverges from that of canonical regulatory elements in the rest of the genome. Specifically, on the same chromatin fibre, these enhancer elements can show the unexpected co-occurrence of chromatin accessibility and hyper-5mCpG methylation, two epigenetic marks widely thought to be mutually exclusive. This plant-specific epigenetic pattern raises questions about the nature of the DNA-binding proteins generating accessibility at these sites, which we show can directly occupy these hyper-5mCpG-methylated accessible elements. Taken together, these results suggest that plant-specific DNA methylation pathways first suppress the proliferation of LTR retrotransposons by silencing their putative promoters, and subsequently prevent the putative LTR enhancers from being exapted as gene enhancers by hyper-5mCpG methylation. In contrast, we present evidence that a primary mechanism by which LTR retrotransposons may shape maize gene regulation is by creating novel gene promoters from ACRs within the LTRs. We also show an example of an ACR-containing LTR retrotransposon facilitating host gene amplification.

Finally, we find that the loci in which Barbara McClintock discovered TE insertions show unusually low gene body methylation and unusually high gene body accessibility, consistent with the common assumption that gene body methylation protects against TE insertion. While the mechanistic underpinnings of these correlated epigenetic features are unclear, this finding adds to our understanding of the complexity of epigenetics and genome evolution in plants.

Methods

Maize mesophyll protoplast generation

We used the polyethylene glycol transformation method of maize mesophyll protoplasts as described in ref. 49. Maize (*Zea mays* L. cultivar B73) seeds were soaked in water overnight at 25 °C. The seeds were germinated in soil for 3 days under long-day conditions (16 h light, 8 h dark) at 25 °C, then moved to complete darkness at 25 °C for 10–11 days. From each seedling, 10-cm sections from the second and third leaf were cut into thin 0.5-mm strips perpendicular to veins and immediately submerged in 10 ml of protoplasting enzyme solution (0.6 M mannitol, 10 mM MES pH 5.7, 15 mg ml⁻¹ cellulase R10, 3 mg ml⁻¹ macerozyme, 1 mM CaCl₂, 0.1% (w/v) BSA and 5 mM beta-mercaptoethanol). The mixture was covered in foil to keep out light, vacuum infiltrated for 3 min

at r.t. and incubated on a shaker at 40 rpm for 2.5 h at r.t. Protoplasts were released by incubating for an extra 10 min at 80 rpm. To quench the reaction, 10 ml ice-cold MMG (0.6 M mannitol, 4 mM MES pH 5.7, 15 mM MgCl₂) was added to the enzyme solution and the whole solution was filtered through a 40-µm cell strainer. To pellet protoplasts, the filtrate was split into equal volumes of no more than 10 ml in chilled round-bottom glass centrifuge vials and centrifuged at 100 × *g* for 4 min at r.t. Pellets were resuspended in 1 ml cold MMG each and combined into a single round-bottom vial. To wash, MMG was added to make a total volume of 5 ml, and the solution was centrifuged at 100 × *g* for 3 min at r.t. This wash step was repeated two more times. The final pellet was resuspended in 1–2 ml of MMG. A sample of the resuspended protoplasts was diluted 1:20 in MMG and used to count the number of viable cells using fluorescein diacetate as a dye.

ATAC-seq data collection

An aliquot of 50,000 isolated protoplasts was added to new tubes and spun down (4 °C, 2,000 *g*) for 10 min. Supernatant was discarded and the pellet of protoplasts was washed with 750 µl of lysis buffer (0.4 M sucrose, 10 mM MgCl₂, 25 mM Tris-HCl pH 8.0, 0.1× protease inhibitor, 0.5% Triton X). Samples were then spun down (4 °C, 1,500 *g*) for 5 min and the supernatant discarded. Samples were then washed once more with buffer (0.4 M sucrose, 10 mM MgCl₂, 25 mM Tris-HCl pH 8.0, 0.1× protease inhibitor) at 4 °C and 1,500 *g* for 3 min to remove the lysis buffer. The nuclear pellet was then resuspended in 22.5 µl double distilled H₂O followed by adding 25 µl of 2× TD buffer (20 mM Tris-HCl pH 7.6, 10 mM MgCl₂, 20% (v/v) dimethylformamide) and 2.5 µl of Tn5. Samples were then incubated at 37 °C for 5 min. Reaction was stopped by adding 250 µl of Zymo Research DNA Binding buffer and DNA was purified using Zymo research Clean and Concentrator kit. Samples were size selected using 1.8× ampure beads and barcoded with Illumina Nextera Index primers. Final library concentrations were determined using Qubit DNA HS assay, and average fragment length was determined using TapeStation D1000 ScreenTape Assay.

Fiber-seq data collection

Isolated protoplasts (1–5 million) were spun down at 2,000 *g* and resuspended in a 100 µl working buffer (400 mM sucrose, 15 mM Tris-Cl, 15 mM NaCl, 60 mM KCl, 1 mM EDTA, 0.5 mM EGTA, 0.5 mM spermidine), with 1.5 µl of 32 mM S-adenosylmethionine added to a final concentration of 0.8 mM along with 0.5 µl of Hia5 MTase (100 U), then carefully mixed by pipetting 10 times with wide bore tips. Reactions were incubated for 10 min at 25 °C, then stopped with 3 µl of 20% SDS (1% final concentration) and transferred to new 1.7 ml microfuge tubes. High molecular weight DNA was then extracted using the Promega Wizard HMW DNA extraction kit A2920. PacBio SMRTbell libraries were then constructed using the manufacturer's SMRTbell prep kit 3.0 procedure. Two replicate samples were processed.

Quantification of 6mA/dA by UHPLC–MS/MS

Samples for quantification were treated as previously described¹⁵. In brief, 50 ng of DNA from each sample was mixed with 0.02 U phosphodiesterase I (Worthington, LS003926), 1 U Benzonase (Millipore Sigma, E1014) and 2.5 U Quick CIP (NEB, M0525S) in digestion buffer (10 mM Tris, 1 mM MgCl pH 8 at r.t.) for 3 h at 37 °C. Single nucleotides were separated from the enzymes by collecting the flow-through of a Nanosep centrifugal filter (MWCO 3 kDa, Pall, OD003C33). The UHPLC–MS/MS analysis of adenosine and m6A was performed on an ACQUITY Premier UPLC System coupled with a XEVO-TQ-XS triple quadrupole mass spectrometer. UPLC was performed on a ZORBAX Eclipse Plus C18 column (2.1 × 50 mm i.d., 1.8 µm particle size) (Agilent, 959757–902). 6mA and dA were eluted and separated using 2–50% linear gradient of solvent B (0.1% acetic acid in 100% methanol) in solvent A (0.1% acetic acid in water) within 10 min, at a flow rate of 0.3 ml min^{−1}. MS/MS analysis was operated in positive ionization mode with 3,000 V

capillary voltage, as well as 150 °C and 1,000 l h^{−1} nitrogen drying gas. A multiple reaction monitoring (MRM) mode was adopted with the following *m/z* transition: 252.10 → 136.09 for dA (collision energy, 14 eV) and 266.2 → 150.2 for 6mA (collision energy, 15 eV). 6mA and dA were mixed to create standards from 0 to 100 nM 6mA, and a new standard was measured and used for each run. MassLynX was used to quantify the data.

Fiber-seq data processing

Fibertools⁵⁰ was used to call m6A methylation and label regions as MSPs and nucleosomes on individual reads. Fiberseq-FIRE was used to assign FDR values to MSPs and call Fiber-seq ACRs using a FIRE model customized for the maize genome⁵¹. Since we trained our model and called FIRE ACRs and elements, the model format used by fiberseq-FIRE has changed. To accommodate these changes, a maize model should be re-trained in the future, which can be readily performed and documented in the fiberseq-FIRE GitHub repository^{17,52}. For ACR calling, we used the set of peaks identified by the FIRE pipeline with an FDR threshold of 1%. Data for two replicates were combined.

ATAC-seq data processing

ATAC-seq read pairs were aligned to the MaizeV5 reference genome¹³ using bwa (v.0.7.17-r118)⁵³. The resulting bam files were filtered using samtools view⁵⁴ to discard reads that were unmapped (−F 4) or had map quality of zero (−q 1). ACRs were called using MACS2 (v.2.2.7.1)⁵⁵, and the narrowPeaks output was merged to generate a non-overlapping set of ACRs. The ATAC-seq signal track is a sliding-window histogram displaying the number of ATAC read ends, with the height of each 20-bp bar representing the number of Tn5 insertions within a 100-bp window centred on that 20 bp.

RNA-seq data used to define expression quantiles and TSSs

Publicly available RNA-seq reads (64,500,061) were obtained from NCBI SRA: [ERR3322830](https://www.ncbi.nlm.nih.gov/sra/ERR3322830). These reads were derived from the second leaves of 9-day-old etiolated seedlings⁵⁶. Reads were aligned to the maizeV5 annotation using hisat2 and counts were tallied using htseq-count. Transcripts per million (TPM) were calculated for each gene. A total of 13,542 genes had a TPM of zero. The remainder were split into deciles by expression level, with each decile containing 2,991 or 2,992 genes. TSS positions were obtained using CAGE data⁵⁷.

Methylation rate (m6A and m5CpG)

For each genomic locus being aggregated, at each 20 bp bin, the number of possible methylation sites was calculated from the individual fibre sequences. The observed methylation events were tallied and divided by the number of possible sites to get a fraction of sites methylated.

MSP score and FIRE accessibility score

MSP score is the fraction of fibre-bases within a given region that are annotated as MSPs. FIRE accessibility score is the fraction of fibre-bases within a given region that are annotated as FIRE element (Extended Data Fig. 1b).

Percent actuation

For a given genomic region, percent actuation is the number of unique reads (fibres) with at least one FIRE element overlapping the region, divided by the total number of unique reads overlapping the region.

Comparing single-cell ATAC-seq to Fiber-seq

The sparse matrix containing binary (cut or no cut) information for all cells from all tissues and all peaks reported in ref. 19 were downloaded from <https://ftp.ncbi.nlm.nih.gov/geo/series/GSE155nnn/GSE155178/suppl/GSE155178%5FACR%5F%5Fcell.binary.sparse.txt.gz>. We then generated a bed file consisting of only peaks with at least one Tn5

insertion in a leaf-designated cell, and reported, for each peak, the fraction of total leaf cells having one or more Tn5 insertions at that site. We used liftOver⁵⁸ to convert the genomic positions of the peaks in this file from V4 to V5 coordinates, then filtered peaks to retain those that (1) overlap by MACS2 peaks by 100 bp or more and (2) overlap our FIRE-peaks by 100 bp or more; 39,132 peaks remain. The percentage of cells with one or more Tn5 insertions was plotted against the percentage of fibres containing a FIRE element (% actuation) in Fig. 1g. For this analysis, 58,712 MACS2 narrowPeaks were called on an alignment file (bam) containing 94,945,002 mapped 50-bp paired-end reads. These peaks were merged (bedops -m), resulting in 50,349 MACS2 peaks used above.

Short-read mappability analysis

A total of 2.1 billion fragments were generated evenly distributed across the B73 reference genome chromosomes 1–10. Fragment lengths were sampled from a log-normal distribution fit to one of our ATAC-seq data sets. For each simulated fragment, a paired-end read was generated with 50 base reads on either end of the fragment. The true start and end of the fragment was encoded in the read name. We did not simulate per-base errors in these reads; each read matches exactly the reference sequence from which it was generated. These reads were then mapped back to the genome using BWA. The ‘fraction mapped’ for a given region or window was calculated as the number of correctly mapped reads with mapq score >0, divided by the total number of simulated reads with the outer end (Tn5 insertion) falling in the region (Extended Data Fig. 2b).

Annotation of repetitive regions, including all TEs

Annotation file Zm-B73-REFERENCE-NAM-5.0.TE.gff3.gz was downloaded from maizedb.org⁵⁹.

Annotation of regions of the nuclear genome with homology to organellar genomes

Regions of homology within the nuclear genome to organellar genomes were identified as follows for each of the chloroplast and mitochondrial genomes, separately. Paired-end reads were simulated to achieve 100× coverage (142,724 and 579,124 read pairs, respectively), then mapped to the MaizeV5 reference genome¹³ using bwa (v.0.7.17-r118)⁵³. The resulting bams were filtered using samtools view⁵⁴ to discard reads that were unmapped (-F 4), had map quality of zero (-q 1), or mapped to the centromere⁶⁰. Alignment files (bams) were then converted to bed files and overlapping regions were merged.

Classification of ACRs

ACRs were sorted hierarchically as follows: (1) all ACRs with low FIRE accessibility score were included in the ‘low FIRE score’ set (medium grey), (2) ACRs with FIRE element (FE) length <200 bp and high FIRE accessibility score were included in the ‘FE length <200’ set (blue), (3) ACRs with mappability <80% and both high FIRE accessibility score and FE length ≥200 bp were included in the ‘Low mappability’ set (dark grey) and (4) ACRs with high FIRE accessibility score, high FE length and high mappability are in the ‘Regular ACRs’ set (light grey).

FiberHMM

Footprints were called on Fiber-seq reads using FiberHMM v.1.3.1. FiberHMM is based on a hidden Markov model (HMM) with two hidden states: accessible and inaccessible. The emission probabilities used in the model are probabilities of methylation of a given base within a 6-nt sequence context in a known accessible or inaccessible state, based on experiment-derived methylation rates from a dechromatinized or a Hia5 untreated control dataset, respectively²³. Transition and starting probabilities for the HMM were trained on a group of 1,000 reads sampled from the Maize dataset shuffled in order 20 times, with initial probabilities picked from the Dirichlet distribution with all parameters set to 1. The best model was chosen and then used for all subsequent footprint calling.

Identifying nearby enhancers

The precise bounds of accessible putative enhancer regions were found using a Gaussian Mixture Model Hidden Markov model (GMM-HMM) to segment regions on the basis of percent nucleosome footprint occupancy. As the positioned mononucleosome footprint found downstream of putative enhancers provided a consistent and clearly defined reference position, the upstream edge of that nucleosome was used to centre the footprint occupancy patterns for metaprofiles of the enhancers.

Labelling footprints in Fiber-seq

Nucleosome footprints were defined as footprints greater than 90 bp. Putative transcription factor footprints were identified on the basis of a combination of their sub-40 bp size and their overlap with known motifs. Transcription factor footprints overlapping multiple motifs were assigned multiple possible identities, indicated in the corresponding visualization. Putative polymerase footprints were defined on the basis of position relative to the predicted TSS of the promoter, and a size of between 40–80 bp, as previously observed in ref. 61.

Enrichment of GWAS SNPs within different classes of ACRs

Single-nucleotide polymorphisms (SNPs) associated with 41 distinct phenotypes⁶² were used to assess whether newly called FIRE ACRs have a similar enrichment of genome-wide association study (GWAS) SNPs to ATAC-called ACRs. GWAS SNPs with reads mapped in peaks <0.05 were removed as described in the paper. FIRE ACRs were split into two categories on the basis of whether they overlap ATAC-seq ACRs as in Fig. 2c. For both categories, an enrichment was calculated by comparing the fraction of ACR bases covered by GWAS SNPs to the fraction covered in the shifted control category. FIRE ACRs overlapping and not overlapping ATAC-seq ACRs were found to have enrichment values of 3.37 and 3.16, respectively.

Calling differential ACRs

ATAC-seq reads from the following six tissues were downloaded from the NCBI Gene Expression Omnibus¹⁹: Tassel (GSM4696882), Ear (GSM4696883), Root1 (GSM4696884), Axillary_bud1 (GSM4696886), Crown_root1 (GSM4696888), Leaf2 (GSM4696890). For each sample, 100 million read pairs were downloaded, trimmed to 50 bp. Each of the six downloaded samples as well as reads from our in-house dark leaf protoplast sample were aligned to the MaizeV5 reference genome¹³ using bwa (v.0.7.17-r118)⁵³. The resulting bams were filtered using samtools view⁵⁴ to discard reads that were unmapped (-F 4), had map quality of zero (-q 1), or mapped to the centromere⁶⁰. Because the number of MACS2 peaks is correlated with the number of mapped reads, for each of the seven samples, the number of aligned reads was subsampled to 16 M. Peaks were called using MACS (v.2.2.7.1)⁵⁵, and the narrowPeaks output was merged to generate a non-overlapping set of peaks for each of the seven samples. A union set of 80,641 peaks was generated by merging the seven sets of peaks (bedops -m)⁶¹. Tn5 insertions were tallied in each union peak for each of the seven samples, and per-bp accessibility was calculated by dividing by the peak length. Because our aim was to find differential ACRs that were inaccessible in dark leaf protoplast, we defined differential ACRs as those that (1) had fewer per-bp Tn5 insertions than twice the minimum dark leaf protoplast cutcounts in a union peak overlapping a called dark leaf protoplast peak and (2) the difference between the accessibility of most-accessible sample and the dark leaf protoplast sample was in the 75th percentile or greater. These 2,826 dACRs are in Supplementary Table 8.

Identification of solo LTRs

LTR sequences from intact LTR retrotransposons containing at least one FIRE ACR within either LTR were aligned to the maize genome using blastn⁶³. Matches with bitscore greater than 1,400 and length

greater than 1,000 bps were retained and merged (bedops -m). Next, we identified matches that (1) did not overlap another intact LTR retrotransposon and (2) contained a FIRE ACR. These are listed in Supplementary Table 10.

Identification of hAT insertion sites

hAT insertion sites are defined as 200-bp windows centred on the location of a hAT transposon polymorphism in which B73 lacks the hAT transposon and exactly 1 of the 25 NAM lines contains a hAT transposon⁴⁵.

Reporting summary

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

Data availability

Raw and processed sequencing data are available from the NCBI Short-Read Archive (SRA) under Bioproject [PRJNA1119563](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1119563).

References

- McClintock, B. The origin and behavior of mutable loci in maize. *Proc. Natl Acad. Sci. USA* **36**, 344–355 (1950).
- McClintock, B. Induction of instability at selected loci in maize. *Genetics* **38**, 579–599 (1953).
- McClintock, B. The significance of responses of the genome to challenge. *Science* **226**, 792–801 (1984).
- Fedoroff, N. V. Transposable genetic elements in maize. *Sci. Am.* **250**, 84–99 (1984).
- Jones, R. N. McClintock's controlling elements: the full story. *Cytogenet. Genome Res.* **109**, 90–103 (2005).
- Fedoroff, N. V. McClintock's challenge in the 21st century. *Proc. Natl Acad. Sci. USA* **109**, 20200–20203 (2012).
- McClintock, B. Controlling elements and the gene. *Cold Spring Harb. Symp. Quant. Biol.* **21**, 197–216 (1956).
- Cui, X. & Cao, X. Epigenetic regulation and functional exaptation of transposable elements in higher plants. *Curr. Opin. Plant Biol.* **21**, 83–88 (2014).
- Makarevitch, I. et al. Transposable elements contribute to activation of maize genes in response to abiotic stress. *PLoS Genet.* **11**, e1004915 (2015).
- Zhao, H. et al. Proliferation of regulatory DNA elements derived from transposable elements in the maize genome. *Plant Physiol.* **176**, 2789–2803 (2018).
- Noshay, J. M. et al. Assessing the regulatory potential of transposable elements using chromatin accessibility profiles of maize transposons. *Genetics* **217**, 1–13 (2021).
- 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).
- Hufford, M. B. et al. De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science* **373**, 655–662 (2021).
- Stergachis, A. B., Debo, B. M., Haugen, E., Churchman, L. S. & Stamatoyannopoulos, J. A. Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science* **368**, 1449–1454 (2020).
- Kong, Y. et al. Critical assessment of DNA adenine methylation in eukaryotes using quantitative deconvolution. *Science* **375**, 515–522 (2022).
- Tullius, T. W. et al. RNA polymerases reshape chromatin and coordinate transcription on individual fibers. *Mol. Cell* <https://doi.org/10.1016/j.molcel.2024.08.013> (2023).
- Vollger, M. R. et al. A haplotype-resolved view of human gene regulation. Preprint at *bioRxiv* <https://doi.org/10.1101/2024.06.14.599122> (2024).
- Bubb, K. L. & Deal, R. B. Considerations in the analysis of plant chromatin accessibility data. *Curr. Opin. Plant Biol.* **54**, 69–78 (2020).
- Marand, A. P., Chen, Z., Gallavotti, A. & Schmitz, R. J. A cis-regulatory atlas in maize at single-cell resolution. *Cell* **184**, 3041–3055.e21 (2021).
- Uren, A. G., Kool, J., Berns, A. & van Lohuizen, M. Retroviral insertional mutagenesis: past, present and future. *Oncogene* **24**, 7656–7672 (2005).
- Erdmann, R. M. & Picard, C. L. RNA-directed DNA methylation. *PLoS Genet.* **16**, e1009034 (2020).
- Jores, T. et al. Synthetic promoter designs enabled by a comprehensive analysis of plant core promoters. *Nat. Plants* **7**, 842–855 (2021).
- Tullius, T. W. et al. RNA polymerases reshape chromatin architecture and couple transcription on individual fibers. *Mol. Cell* **84**, 3209–3222.e5 (2024).
- Leduque, B., Edera, A., Vitte, C. & Quadrana, L. Simultaneous profiling of chromatin accessibility and DNA methylation in complete plant genomes using long-read sequencing. *Nucleic Acids Res.* **52**, 6285–6297 (2024).
- Colonna Romano, N. & Fanti, L. Transposable elements: major players in shaping genomic and evolutionary patterns. *Cells* **11**, 1048 (2022).
- Cohen, C. J., Lock, W. M. & Mager, D. L. Endogenous retroviral LTRs as promoters for human genes: a critical assessment. *Gene* **448**, 105–114 (2009).
- Sundaram, V. et al. Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res.* **24**, 1963–1976 (2014).
- Medstrand, P., Landry, J. R. & Mager, D. L. Long terminal repeats are used as alternative promoters for the endothelin B receptor and apolipoprotein C-I genes in humans. *J. Biol. Chem.* **276**, 1896–1903 (2001).
- Dunn, C. A., Medstrand, P. & Mager, D. L. An endogenous retroviral long terminal repeat is the dominant promoter for human beta1,3-galactosyltransferase 5 in the colon. *Proc. Natl Acad. Sci. USA* **100**, 12841–12846 (2003).
- Stelpflug, S. C. et al. An expanded maize gene expression atlas based on RNA sequencing and its use to explore root development. *Plant Genome* <https://doi.org/10.3835/plantgenome2015.04.0025> (2016).
- Yang, N. et al. Two teosintes made modern maize. *Science* **382**, eadg8940 (2023).
- Bailey, J. A., Liu, G. & Eichler, E. E. An Alu transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genet.* **73**, 823–834 (2003).
- Cao, J. et al. Epigenetic and chromosomal features drive transposon insertion in *Drosophila melanogaster*. *Nucleic Acids Res.* **51**, 2066–2086 (2023).
- 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).
- Fedoroff, N., Wessler, S. & Shure, M. Isolation of the transposable maize controlling elements Ac and Ds. *Cell* **35**, 235–242 (1983).
- Courage-Tebbe, U., Döring, H. P., Fedoroff, N. & Starlinger, P. The controlling element *Ds* at the *Shrunken* locus in *Zea mays*: structure of the unstable *sh-m5933* allele and several revertants. *Cell* **34**, 383–393 (1983).
- Shure, M., Wessler, S. & Fedoroff, N. Molecular identification and isolation of the *Waxy* locus in maize. *Cell* **35**, 225–233 (1983).

38. Fedoroff, N. V., Furtek, D. B. & Nelson, O. E. Cloning of the *bronze* locus in maize by a simple and generalizable procedure using the transposable controlling element *Activator* (Ac). *Proc. Natl Acad. Sci. USA* **81**, 3825–3829 (1984).
39. Paz-Ares, J., Wienand, U., Peterson, P. A. & Saedler, H. Molecular cloning of the *c* locus of *Zea mays*: a locus regulating the anthocyanin pathway. *EMBO J.* **5**, 829–833 (1986).
40. Elgin, S. C. DNAase I-hypersensitive sites of chromatin. *Cell* **27**, 413–415 (1981).
41. Sullivan, A. M. et al. Mapping and dynamics of regulatory DNA and transcription factor networks in *A. thaliana*. *Cell Rep.* **8**, 2015–2030 (2014).
42. French, S. L., Osheim, Y. N., Cioci, F., Nomura, M. & Beyer, A. L. In exponentially growing *Saccharomyces cerevisiae* cells, rRNA synthesis is determined by the summed RNA polymerase I loading rate rather than by the number of active genes. *Mol. Cell. Biol.* **23**, 1558–1568 (2003).
43. Takuno, S. & Gaut, B. S. Body-methylated genes in *Arabidopsis thaliana* are functionally important and evolve slowly. *Mol. Biol. Evol.* **29**, 219–227 (2012).
44. Muyle, A. M., Seymour, D. K., Lv, Y., Huettel, B. & Gaut, B. S. Gene body methylation in plants: mechanisms, functions, and important implications for understanding evolutionary processes. *Genome Biol. Evol.* **14**, evac038 (2022).
45. Munasinghe, M. et al. Combined analysis of transposable elements and structural variation in maize genomes reveals genome contraction outpaces expansion. *PLoS Genet.* **19**, e1011086 (2023).
46. Fedoroff, N. V. Transposable elements, epigenetics, and genome evolution. *Science* **338**, 758–767 (2012).
47. Mo, W. et al. Single-molecule targeted accessibility and methylation sequencing of centromeres, telomeres and rDNAs in *Arabidopsis*. *Nat. Plants* **9**, 1439–1450 (2023).
48. Jha, A. et al. DNA-m6A calling and integrated long-read epigenetic and genetic analysis with fibertools. *Genome Res.* <https://doi.org/10.1101/gr.279095.124> (2024).
49. Tonnies, J., Mueth, N. A., Gorjifard, S., Chu, J. & Queitsch, C. Scalable transfection of maize mesophyll protoplasts. *J. Vis. Exp.* <https://doi.org/10.3791/64991> (2023).
50. Fibertools-Rs: tools for fiberseq data written in rust. *GitHub* <https://github.com/fiberseq/fibertools-rs> (2024).
51. Hamm, M. Fiber-seq FIRE maize model. *Zenodo* <https://doi.org/10.5281/ZENODO.14641792> (2025).
52. FIRE: a snakemake workflow for calling fiber-seq inferred regulatory elements (FIREs) on single molecules. *GitHub* <https://github.com/fiberseq/FIRE> (2024).
53. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
54. SAMtools. *GitHub* <https://samtools.sourceforge.net/> (2024).
55. Gaspar, J. M. Improved peak-calling with MACS2. Preprint at *bioRxiv* <https://doi.org/10.1101/496521> (2018).
56. Hendron, R.-W. & Kelly, S. Subdivision of light signaling networks contributes to partitioning of C4 photosynthesis. *Plant Physiol.* **182**, 1297–1309 (2020).
57. Mejía-Guerra, M. K. et al. Core promoter plasticity between maize tissues and genotypes contrasts with predominance of sharp transcription initiation sites. *Plant Cell* **27**, 3309–3320 (2015).
58. Hinrichs, A. S. et al. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006).
59. Welcome to MaizeGDB (MaizeGDB, 2024); <https://www.maizegdb.org/>
60. Wolfgruber, T. K. et al. Maize centromere structure and evolution: sequence analysis of centromeres 2 and 5 reveals dynamic loci shaped primarily by retrotransposons. *PLoS Genet.* **5**, e1000743 (2009).
61. Neph, S. et al. BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**, 1919–1920 (2012).
62. Wallace, J. G. et al. Association mapping across numerous traits reveals patterns of functional variation in maize. *PLoS Genet.* **10**, e1004845 (2014).
63. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
64. Pb-CpG-Tools: collection of tools for the analysis of CpG data. *GitHub* <https://github.com/PacificBiosciences/pb-CpG-tools> (2024).

Acknowledgements

We thank S. Fields, O. Waltner, M. Stitzer, E. Buckler and J. Ross-Ibarra for helpful data analysis suggestions and discussion of results, as well as detailed manuscript comments. This work was supported by the National Science Foundation (PlantSynBio Grant No. 2240888 to C.Q., NSF Postdoctoral Research Fellowship in the Biology Program Grant No. 2305660 to B.R.-C.), the National Institutes of Health (NIGMS MIRA Grant No. 1R35GM139532 to C.Q.), and the United States Department of Agriculture (NIFA postdoctoral fellowship no. 2023-67012-39445 to N.A.M.). A.B.S. holds a Career Award for Medical Scientists from the Burroughs Wellcome Fund and is a Pew Biomedical Scholar. This study was supported by National Institutes of Health (NIH) grants 1DP5OD029630 and UM1DA058220 to A.B.S. M.R.V. was supported by a training grant (T32) from the NIH (2T32GM007454-46).

Author contributions

M.O.H., N.A.M., B.R.-C., Y.M., E.J.B. and J.R. performed experiments, M.O.H., K.L.B., J.K.M., M.R.V. and T.W.T. performed computational analyses. M.O.H., K.L.B., T.W.T., J.K.M., J.T.C., C.T., C.Q. and A.B.S. prepared figures and wrote the paper. M.O.H., J.T.C., C.Q., C.T. and A.B.S. conceived and designed the experiments.

Competing interests

A.B.S. is a co-inventor on a patent relating to the Fiber-seq method (US17/995,058). The other authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41477-025-02002-z>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41477-025-02002-z>.

Correspondence and requests for materials should be addressed to Andrew B. Stergachis or Christine Queitsch.

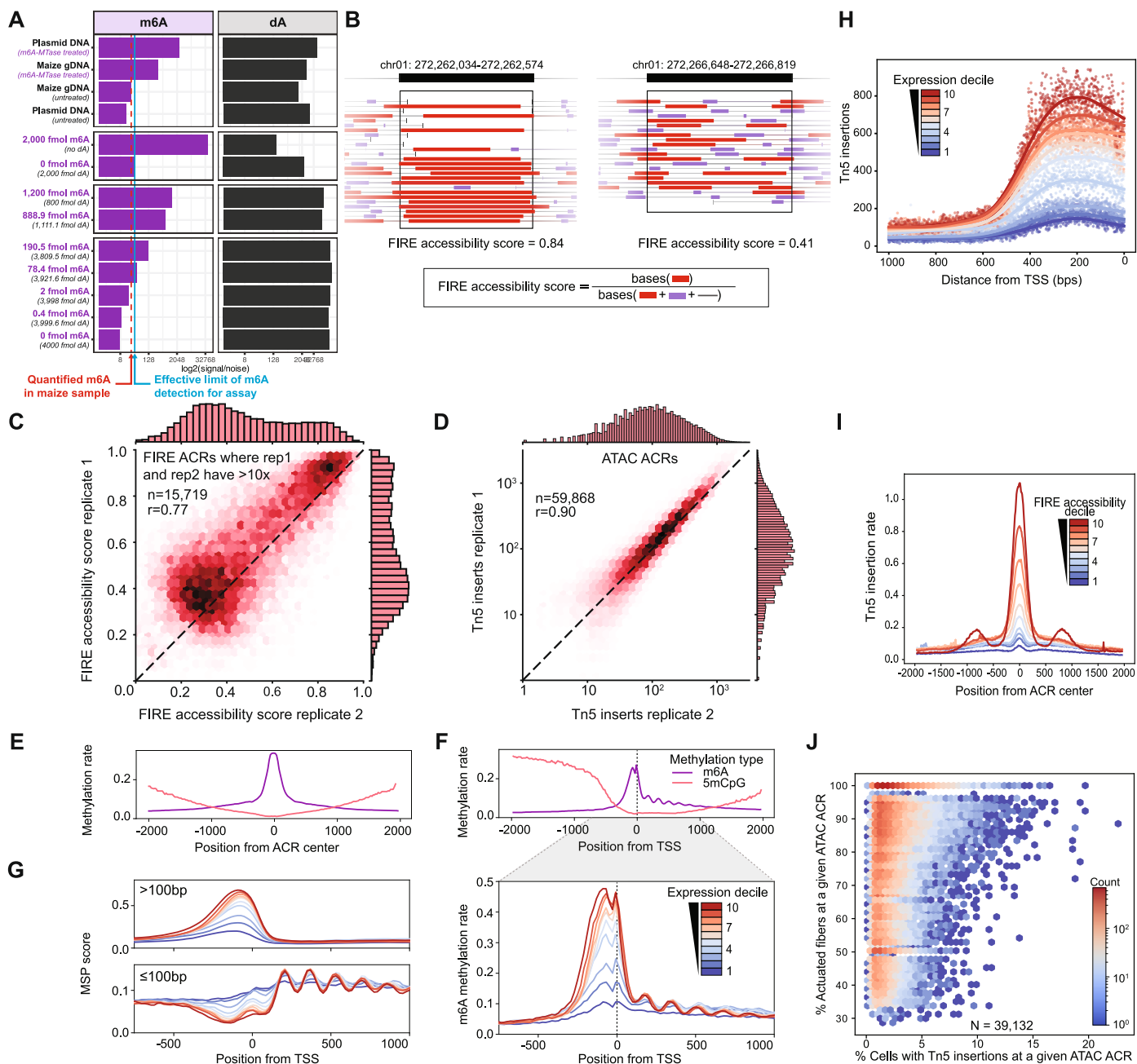
Peer review information *Nature Plants* thanks Jungnam Cho and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

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.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2025

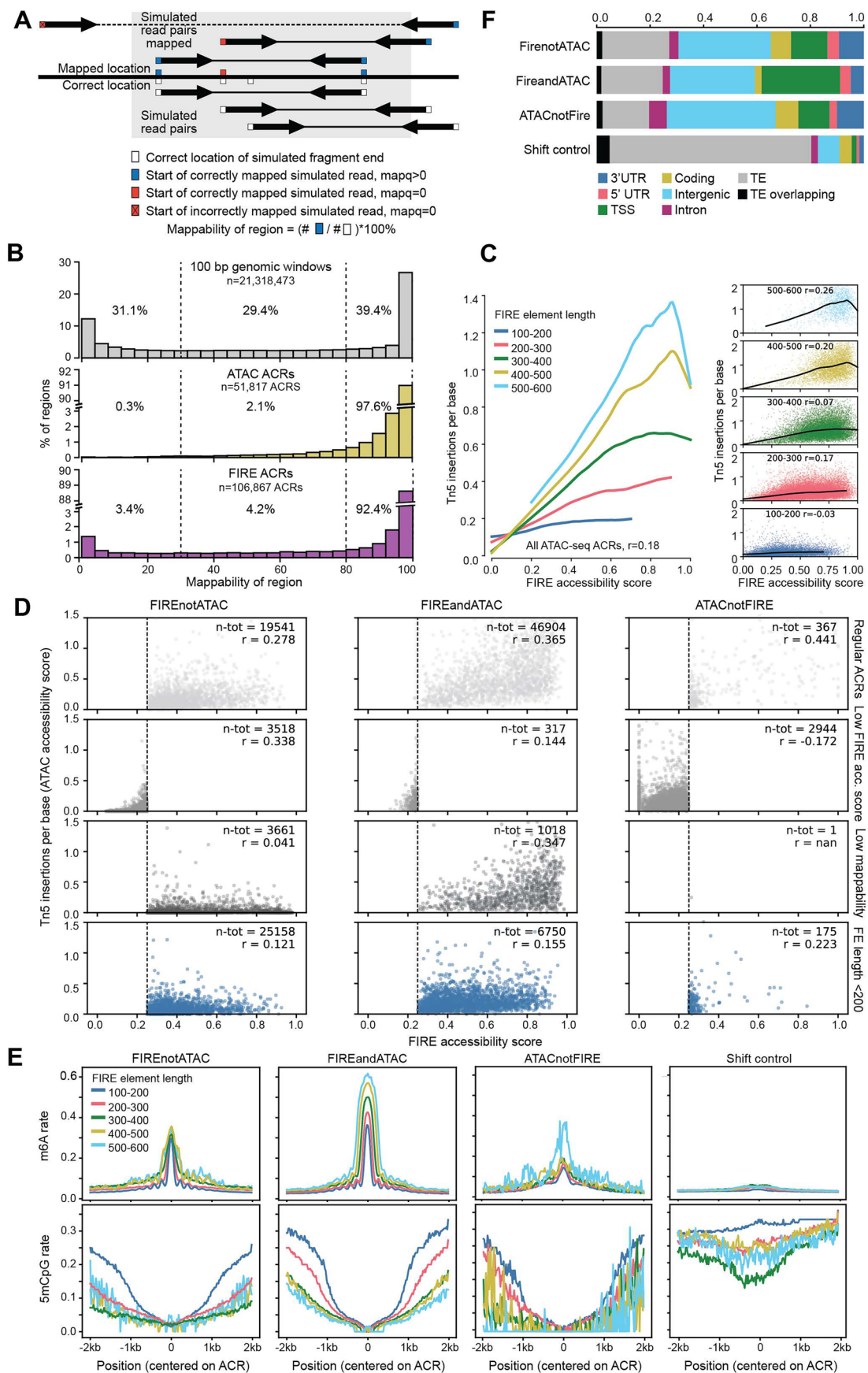


Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Fiber-seq-derived ACRs show expected patterns at ATAC ACRs and transcription start sites and expected correlation with expression.

(A) Small molecule mass spectrometry data on nucleotides purified from various samples with known and unknown quantities of m6A and adenine (dA). Shown is the log-scale of the signal-to-noise ratio for each sample. Specifically, this includes plasmid DNA isolated from bacteria that lack any m6A-MTases. As positive controls these samples were treated with a non-specific m6A-MTase before nucleotide isolation. For maize, samples were prepared as for Fiber-seq with or without m6A-MTase-treatment. In addition, standards containing defined amounts of m6A and dA were used. The red dashed line corresponds to the signal observed in untreated maize genomic DNA, whereas the solid blue line shows essentially the limit of detection of m6A signal-to-noise for this assay based on the sample that has no m6A injected with a similar amount of total dA. **(B)** Schematic illustrating the calculation of the FIRE accessibility score, a measure of Fiber-seq-derived chromatin accessibility that allows direct comparisons to ATAC-seq-derived chromatin accessibility. Shown are screenshots of two FIRE ACRs with individual fibers showing FIRE elements (red) of different length, in addition to methyltransferase-sensitive patches in purple and non-methylated regions as grey lines. Black boxes mark the respective FIRE ACRs (black bars on top). For any given window, the FIRE accessibility score is calculated as the number of bases annotated as FIRE elements (red) divided by the total number of bases across all fibers mapping within this window (red, purple for methyltransferase-sensitive patches not annotated as FIRE elements, grey for not methylated). FIRE accessibility scores are shown for the two example ACRs. **(C)** Correlation between FIRE accessibility scores for Fiber-seq replicates 1 and 2. Each dot corresponds to ACRs where both replicates have >10x coverage. **(D)** Correlation of Tn5 insertions for union ACRs identified in ATAC-seq replicates 1 and 2. **(E)** m6A methylation peaked at the centre of ATAC-seq derived ACRs in paired samples. **(F)** m6A methylation rate peaked

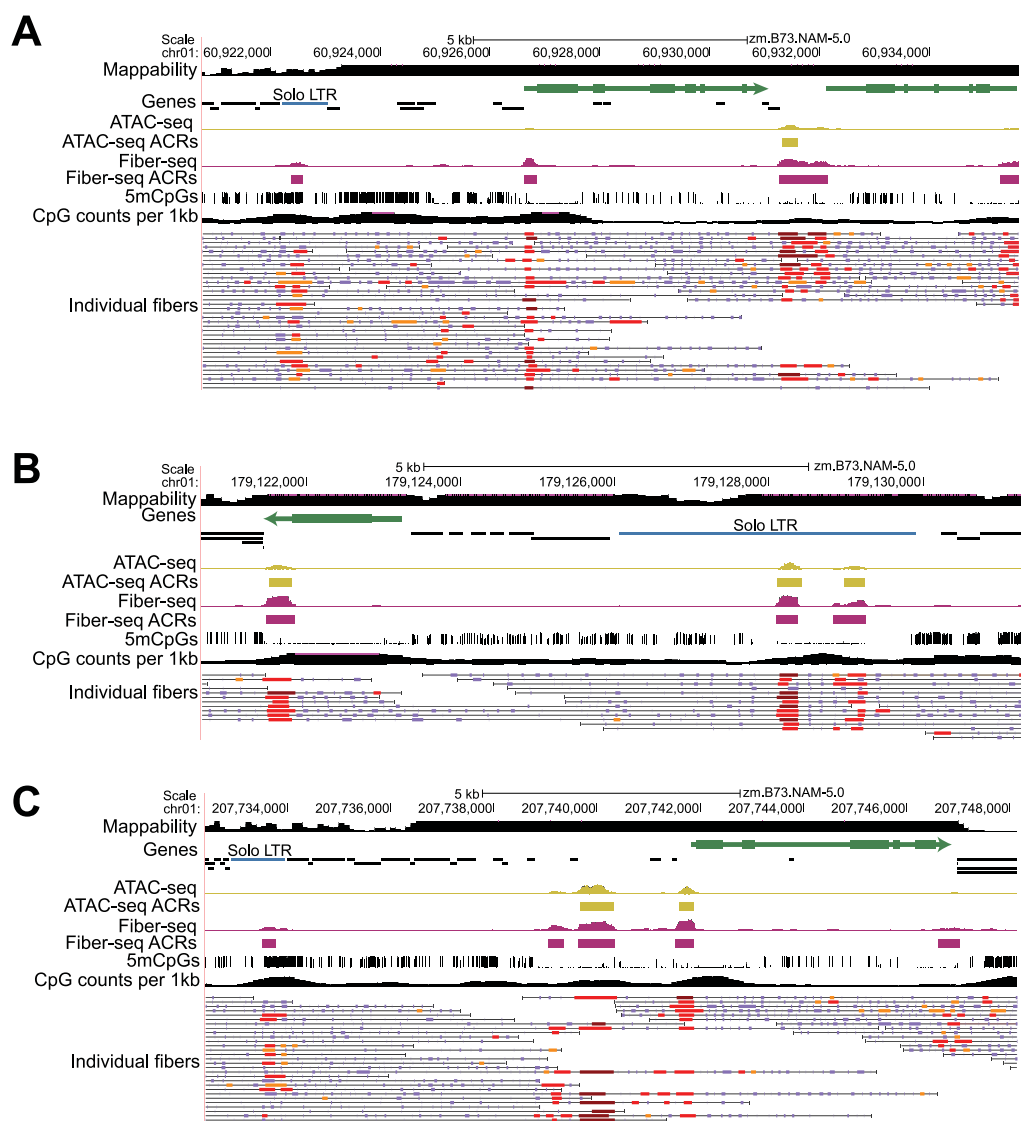
immediately upstream of CAGE-defined transcription start sites (TSSs), with phased nucleosomes apparent downstream of TSSs. Average strength of m6A methylation rate upstream of TSSs was monotonically related to expression level of respective downstream genes (expression deciles). The well-phased-nucleosome signal was strongest for highly expressed genes and faded for lowly expressed ones, as expected. **(G)** Methyltransferase-sensitive patches (MSPs) larger than 100 bp constituted the majority of the m6A signal at TSSs, while MSPs shorter than 100 bp showed patterns consistent with well-positioned nucleosomes. MSP scores were calculated in aggregate for each non-overlapping 20 bp window in the region 750 bp upstream and 1 kb downstream of each TSS (see Methods). **(H)** Aggregate plot of Tn5 insertions/base in the 1 kb window upstream of TSSs stratified by downstream gene expression for paired ATAC-seq data, comparable to **(F)**. (FIRE elements supporting FIRE ACRs are in red, purple indicates methylation sensitive patches (see Fig. 1). **(I)** Aggregate plot of FIRE ACRs stratified into ten deciles based on their FIRE accessibility score. For each FIRE score accessibility decile, the number of Tn5 insertions at each bp within 2 kb of the FIRE ACR centre is shown. Accessibility measured by ATAC-seq and Fiber-seq is monotonically correlated. Highly accessible FIRE ACRs tend to show neighbouring FIRE ACRs (symmetric signal at highest decile). This signal is in part due to FIRE ACRs in low-mappability LTR retrotransposons (see Fig. 2). **(J)** The single-molecule method Fiber-seq outperforms single-cell ATAC-seq as a quantitative measure of chromatin accessibility. 39,132 ACRs were identified as shared FIRE ACRs in dark-grown maize leaves and ATAC ACRs in a pseudobulked leaf sample (GSM4696890) from Marand et al. The percentage of cells containing at least one Tn5 insertion within a shared ACR (% cells accessible) is compared to the percentage of actuated fibers (*that is*, with a called FIRE element, % actuated Fibers within a given ATAC ACR) underlying the same shared ACR. Each dot represents one shared ACR. Hexbin color reflects the number of dots.



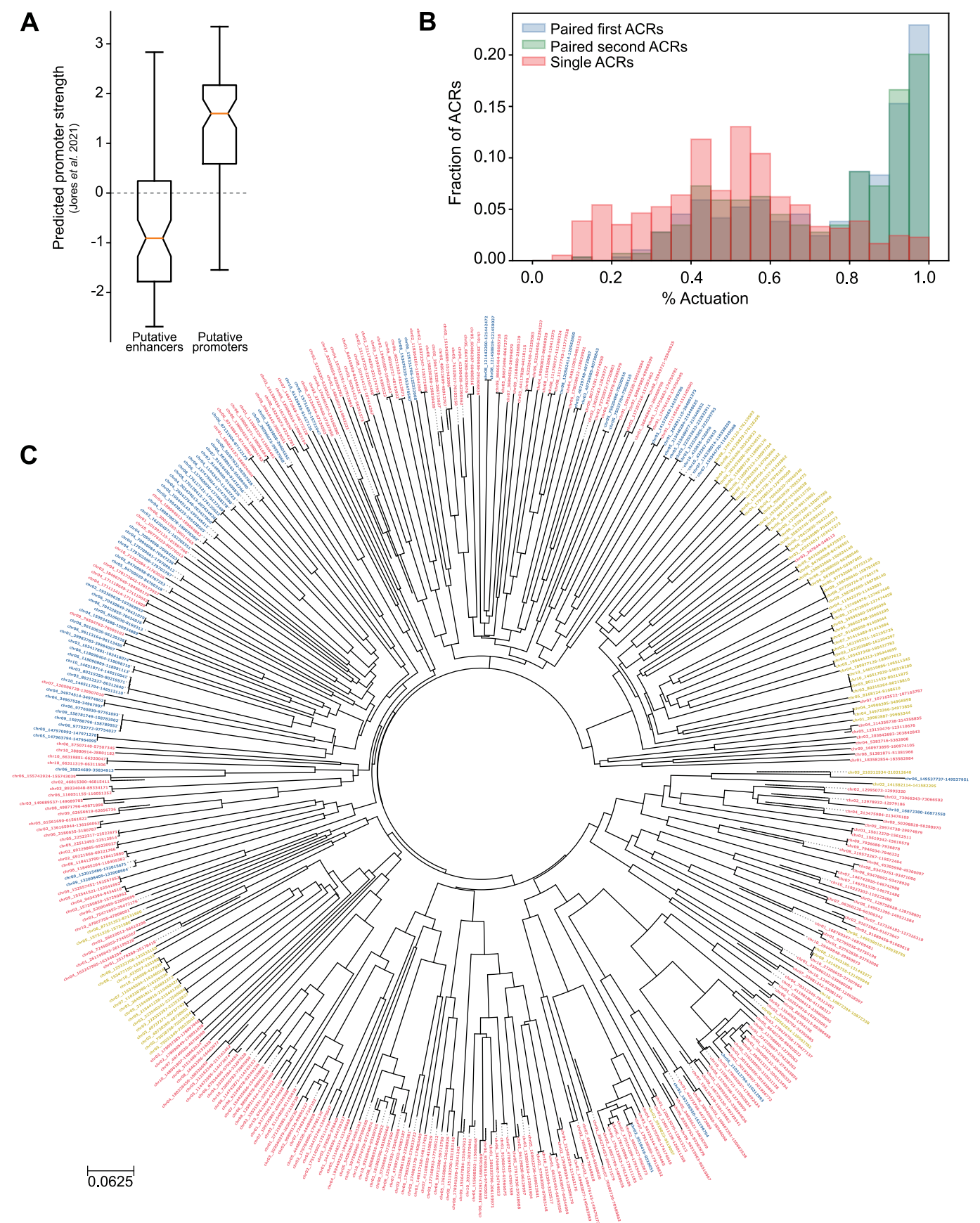
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Novel FIRE ACRs comprised of short FIRE elements are bona fide regulatory elements. (A) Schematic describing short-read simulation and mappability calculation. We generated 2.1 billion fragments evenly distributed across the B73 reference genome chromosomes 1-10 (see Methods). For each simulated fragment, 50 bp paired-end reads were generated (indicated with thick black arrows). Each read matched exactly the reference sequence from which it was generated. These simulated reads were then mapped back to the genome using BWA. The ‘fraction mapped’ for a given region or window was calculated as the number of correctly mapped reads with mapq score > 0 divided by the total number of simulated reads with the outer end (Tn5 insertion) falling in the region. Mapq scores are indicated by blue and red boxes, incorrectly mapped simulated read shows X in red box (top row). Mappability of regions was determined as percentage of correctly mapped reads with mapq>0. (B) Histograms of mappability as in (A) for all 21,318,473 non-overlapping 100 bp windows in the maize genome (top panel, grey), 51,817 ATAC ACRs (middle panel, gold), and 106,867 FIRE ACRs (bottom panel, purple). Low mappability explains only in part why Fiber-seq detects many more ACRs than ATAC-seq. (C) FIRE ACRs comprised of short FIRE elements are not detected by ATAC-seq. Correlation between FIRE accessibility scores and Tn5 insertions/

base (chromatin accessibility as measured by ATAC-seq) for FIRE ACRs comprised of FIRE elements of indicated length (see inset for legend). **Left**, LOWESS curves fitted to FIRE ACRs in respective length categories. **Right**, plots showing individual values for FIRE ACRs belonging to the five length categories. (D) FIRE accessibility score by Tn5 insertions/base (*that is*, ATAC accessibility score) for ACRs stratified into 12 categories. Each dot represents an ACR with the labelled row and column properties. As the row categories are overlapping, ACRs were sorted hierarchically as follows: all ACRs with low FIRE accessibility score were included in the ‘low FIRE acc. score’ rows; ACRs with FE length < 200 bp and high FIRE accessibility score were included in the ‘FE length <200’ rows; ACRs with mappability < 80% and both high FIRE accessibility score and FE length ≥ 200 bp were included in the ‘Unmappable’ rows. (E) FIRE ACRs that do not overlap with ATAC ACRs show similar patterns of the m6A signal (top) and the 5mCpG signal (bottom) as FIRE ACRs that overlap with ATAC ACRs. Shifted control regions do not display these properties. FIRE element length underlying FIRE ACRs is indicated as in (C). (F) FIRE ACRs that do not overlap with ATAC ACRs show a similar distribution across genomic compartments as FIRE ACRs that overlap with ATAC ACRs. Statistical analyses and p-values for Extended Data Fig. 2F are in Supplementary Table 1.



Extended Data Fig. 3 | Examples of solo LTRs containing FIRE ACRs. (A-C) Solo LTRs containing FIRE ACR are colored blue. **(A)** [chr01:60,920,594-60,935,475] **(B)** [chr01:179,120,635-179,131,399] **(C)** [chr01:207,732,409-207,748,141]. See Supplementary Table 10 for a comprehensive list.

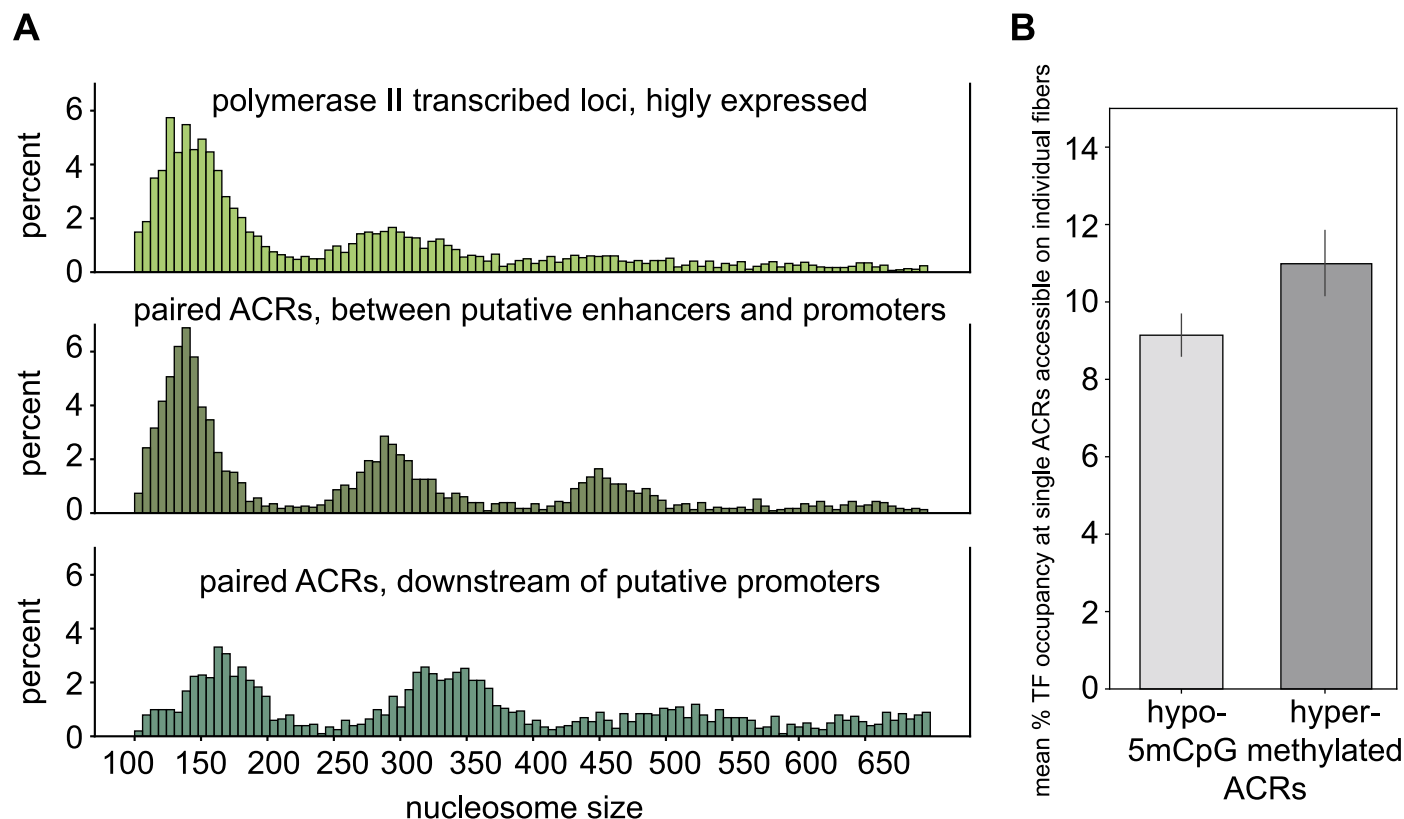


Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Features of FIRE ACRs within LTR retrotransposons.

(A) For each putative enhancer-promoter pair, a sequence starting at the 5' end of the putative enhancer ACR and ending at the 3' end of the putative promoter ACR was extracted. The boxplot shows the predicted promoter strength of the first (coinciding with the putative enhancer) and last 170 bp (coinciding with the putative promoter) of this window. Predictions were made with a CNN model trained on Plant-STARR-seq data for ~75,000 TSS-proximal ACRs (170 bp in length) from Arabidopsis, maize, and sorghum. (B) Histograms for the

percentage actuation (*that is*, the percentage of fibers with a FIRE element that comprise a FIRE ACR) for the first of two paired ACRs (putative enhancers), the second of two paired ACRs (putative promoters), and single ACRs. (C) Phylogeny of LTR ACRs. Branch length units are in estimated substitutions per site. Colors indicate ACR types with blue denoting paired first ACRs (putative enhancers), yellow denoting paired second ACRs (putative promoters) and red denoting single ACRs in LTRs.

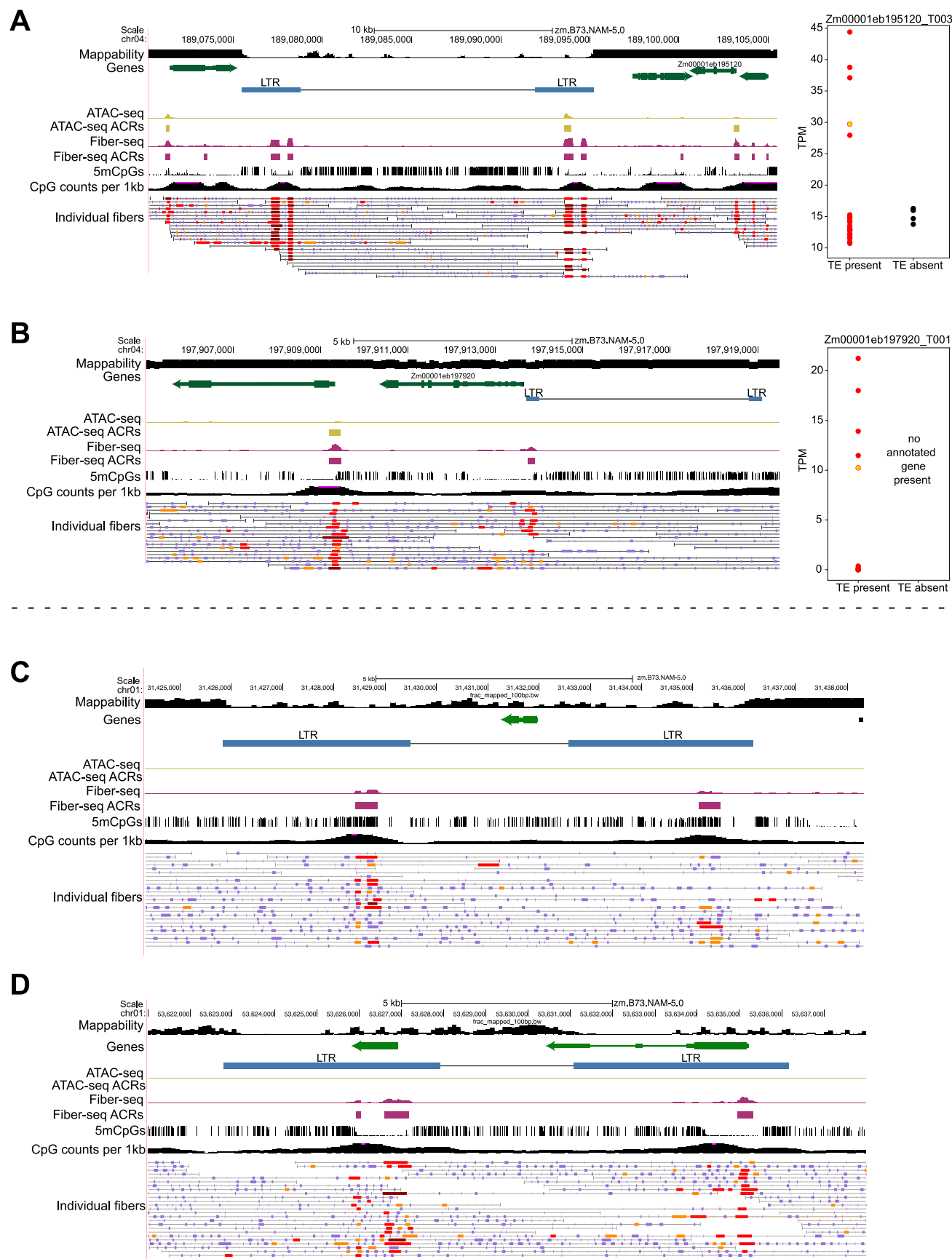


Extended Data Fig. 5 | Nucleosome and TF footprint features of LTR FIRE ACRs.

A) Histogram showing the percent of footprints sized between 100 and 700 bp identified (top) 100 to 1100 bp downstream of the TSS of Pol II genes in the top 3 deciles of expression, (middle) between putative enhancer/promoter pairs, and (bottom) 300–1300 bp downstream of the putative promoter.

(B) Bar plot showing the mean percent putative TF (10–40 bp) occupancy

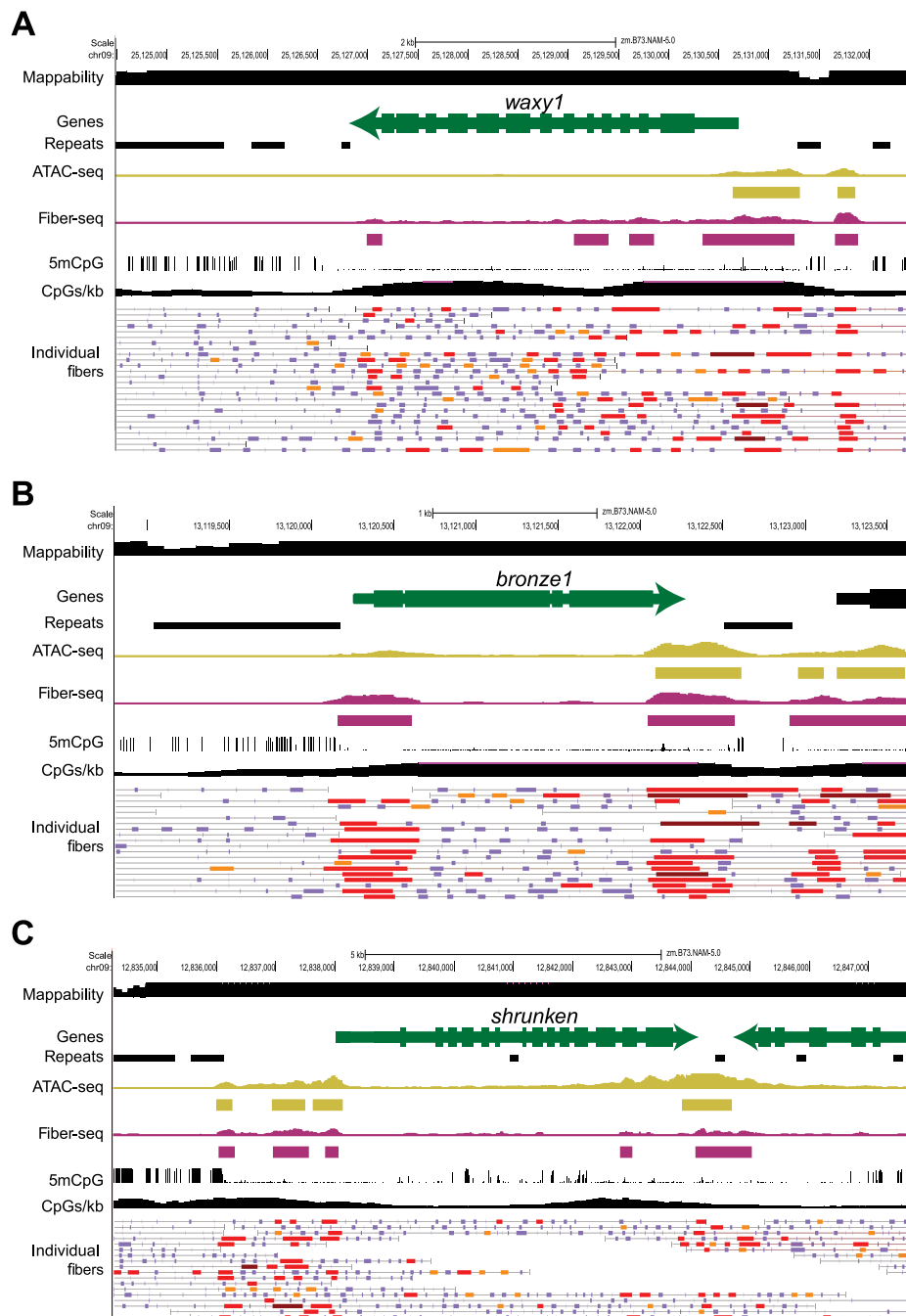
within ± 100 bp of the center of all hypo- or hyper-methylated single ACRs calculated from reads where the center position of the ACR was not occluded by a nucleosome. Error bars represent the 95th percentile range calculated from 10,000x bootstrapped resampling of each group of reads. Statistical analysis and p-value for Extended Data Fig. 5B are in Supplementary Table 1.



Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Examples of ACRs in intact polymorphic LTR retrotransposons (A, B) and LTR retrotransposons with non-TE internal genes (C, D). (A) Left, intact LTR retrotransposon with blue LTRs is absent in NAM lines: Il14H, Ki3, M37W, P39. Tracks in screenshot as in Fig. 2. Right, expression level of indicated gene in lines with and without the TE, B73 is labeled in yellow. (B) Left, intact retrotransposon with blue LTRs is absent in NAM lines: B97, CML228,

CML52, Ki11, Ky21, Mo18W, P39. Tracks in screenshot as in Fig. 2. Right, expression level of indicated gene in lines with and without the TE, B73 is labeled in yellow. (C) Example of an intact LTR retrotransposon containing one annotated gene between the LTRs and lacking an ACR at the transcription start site. (D) Example of an intact LTR retrotransposon containing two annotated genes. For each gene, transcription begins at a FIRE ACR within the LTR.



Extended Data Fig. 7 | Diffuse chromatin accessibility and hypo-5mCpG methylation is observed at loci discovered as hAT TE insertion sites by McClintock. (A) *waxy1* (Zm00001eb378140; chr09:25,127,146 - 25,129,800), one of the first genes identified by McClintock as having a hAT TE insertion, shows higher gene-body chromatin accessibility than 84.4% of other genes. McClintock identified alleles *Ds wx-m9*, *Ds wx-m6*, *Ac wx-m9*, with the *Ds* or *Ac* prefix indicating whether it was a nonautonomous or autonomous hAT TE, respectively. **(B)** *bronze1* (Zm00001eb374230; chr09:13,118,806 - 13,123,664), one of the first genes identified by McClintock as having a hAT TE insertion. McClintock

identified the *Ac bz-m2* allele. The *Ac* prefix indicates insertion of an autonomous hAT TE. **(C)** *shrunk* (Zm00001eb374090; chr09:12,836,508 - 12,845,499), one of the first genes identified by McClintock as having a hAT TE insertion. McClintock identified two germinally-stable alleles, *Ds-4864A* and *Ds-5245*, that were “genetically indistinguishable and located just distal to the Shrunken (*Sh*) locus on the short arm of chromosome 9” and three germinally-unstable alleles, *sh-m6233*, *sh-m5933*, *sh-m6258*, that contain rearrangements at the *Sh* locus related to a hAT insertion, one of which contains a *Ds*-mediated 30 kb insertion. The *Ds* prefix indicates insertion of a nonautonomous hAT TE.

Reporting Summary

Nature Portfolio 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 Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).
Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
[For final submission](#): please carefully check your responses for accuracy; you will not be able to make changes later.

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 (<i>n</i>) 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 <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
☒	☐ 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. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted <i>Give P values as exact values whenever suitable.</i>
☒	☐ 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 <i>d</i> , Pearson's <i>r</i>), 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	bwa v0.7.17-r118, samtools v1.19, bedops v2.4.41, bedtools v2.31.1, macs v2.2.7.1, hisat v2.2.1, htseq v0.12.4, pb-CpG-tools v3.0.0, fiberseq-smk- accessed Jan. 17, 2023, FIRE v0.0.4, FiberHMM v1.3.1

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 and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Raw and processed sequencing data produced for this manuscript (FIBER-seq and ATAC-seq) is available from the NCBI Short Read Archive (SRA) under Bioproject PRJNA1119563. Raw RNA-seq reads were downloaded from NCBI SRA: ERR3322830. The single-cell ATAC-seq data were downloaded from the NCBI Gene Expression Omnibus (Marand et al 2021): GSE155178_ACR_x_cell.binary.sparse.txt.gz (GSE155178), Tassel (GSM4696882), Ear (GSM4696883), GSM4696884 (Root1), Axillary_bud1 (GSM4696886), Crown_root1 (GSM4696888), Leaf2 (GSM4696890).

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

Reporting on race, ethnicity, or other socially relevant groupings

Population characteristics

Recruitment

Ethics oversight

Note that full information on the approval of the study protocol must also be provided in the manuscript.

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](https://www.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

This study was not a case-control study so no power analysis was required. If we interpret n as the depth of coverage of Fiber-seq reads in our study, it was sufficient to achieve greater sensitivity than ATAC-seq, capturing over double the number of ATAC-seq ACRs, and overlapping with more than 90% of those ATAC-seq ACRs.

Data exclusions

Replication

See External Data Fig. 1C,D and corresponding legend for a description of replicates.

Randomization

This study was not an interventional study, so there was no need to randomize test subjects.

Blinding

Behavioural & social sciences study design

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

Study description

Research sample

Sampling strategy

Data collection

Timing

Data exclusions

Non-participation

Randomization

Ecological, evolutionary & environmental sciences study design

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

Study description	
Research sample	
Sampling strategy	
Data collection	
Timing and spatial scale	
Data exclusions	
Reproducibility	
Randomization	
Blinding	

Did the study involve field work? ☐ Yes ☐ No

Field work, collection and transport

Field conditions	
Location	
Access & import/export	
Disturbance	

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		Methods	
n/a	Involved in the study	n/a	Involved in the study
☐	☐ Antibodies	☐	☐ ChIP-seq
☐	☐ Eukaryotic cell lines	☐	☐ Flow cytometry
☐	☐ Palaeontology and archaeology	☐	☐ MRI-based neuroimaging
☐	☐ Animals and other organisms		
☐	☐ Clinical data		
☐	☐ Dual use research of concern		
☐	☐ Plants		

Antibodies

Antibodies used	
Validation	

Eukaryotic cell lines

Policy information about [cell lines and Sex and Gender in Research](#)

Cell line source(s)

Authentication

Mycoplasma contamination

Commonly misidentified lines
(See [ICLAC](#) register)

Palaeontology and Archaeology

Specimen provenance

Specimen deposition

Dating methods

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research, and [Sex and Gender in Research](#)

Laboratory animals

Wild animals

Reporting on sex

Field-collected samples

Ethics oversight

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration

Study protocol

Data collection

Outcomes

Dual use research of concern

Policy information about [dual use research of concern](#)

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes
☐	☐
☐	☐
☐	☐
☐	☐
☐	☐

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes
☐	☐
☐	☐
☐	☐
☐	☐
☐	☐
☐	☐
☐	☐
☐	☐
☐	☐

Plants

Seed stocks	
Novel plant genotypes	
Authentication	

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 <i>May remain private before publication.</i>	
Files in database submission	
Genome browser session (e.g. UCSC)	

Methodology

Replicates	
Sequencing depth	
Antibodies	
Peak calling parameters	
Data quality	

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

Instrument

Software

Cell population abundance

Gating strategy

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

Magnetic resonance imaging

Experimental design

Design type

Design specifications

Behavioral performance measures

Imaging type(s)

Field strength

Sequence & imaging parameters

Area of acquisition

Diffusion MRI

☐ Used

☐ Not used

Preprocessing

Preprocessing software

Normalization

Normalization template

Noise and artifact removal

Volume censoring

Statistical modeling & inference

Model type and settings

Effect(s) tested

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference

(See [Eklund et al. 2016](#))

Correction

Models & analysis

- | | |
|--------------------------|---|
| n/a | Involvement in the study |
| ☐ | ☐ Functional and/or effective connectivity |
| ☐ | ☐ Graph analysis |
| ☐ | ☐ Multivariate modeling or predictive analysis |

Functional and/or effective connectivity

Graph analysis

Multivariate modeling and predictive analysis

