OXFORD GENETICS

GENETICS, 2021, 217(1), iyaa003
DOI: 10.1093/genetics/iyaa003
Advance Access Publication Date: 20 November 2020
Investigation

Assessing the regulatory potential of transposable elements using chromatin accessibility profiles of maize transposons

Jaclyn M. Noshay,[1] Alexandre P. Marand,[2] Sarah N. Anderson,[3] Peng Zhou,[1] Maria Katherine Mejia Guerra,[4] Zefu Lu,[2] Christine H. O'Connor,[5] Peter A. Crisp,[6] Candice N. Hirsch,[5] Robert J. Schmitz,[2] and Nathan M. Springer[1,*]

[1]Department of Plant and Microbial Biology, University of Minnesota, 140 Gortner Laboratory, 1479 Gortner Avenue, St. Paul, MN 55108, USA
[2]Department of Genetics, University of Georgia, 120 W Green St, Athens, GA 30602, USA
[3]Department of Genetics, Development, and Cell Biology, Iowa State University, 2437 Pammel Dr, Ames, IA 50011, USA
[4]Department of Plant Breeding and Genetics, Cornell University, 233 Emerson Hall, Ithaca, NY 14850, USA
[5]Department of Agronomy and Plant Genetics, University of Minnesota, 1994 Upper Buford Circle, 411 Borlaug Hall, St. Paul, MN 55108, USA
[6]School of Agriculture and Food Sciences, The University of Queensland, Harley Teakle Building, Keyhold Rd, St Lucia QLD 4067, Australia

*Corresponding author: springer@umn.edu

Abstract

Transposable elements (TEs) have the potential to create regulatory variation both through the disruption of existing DNA regulatory elements and through the creation of novel DNA regulatory elements. In a species with a large genome, such as maize, many TEs interspersed with genes create opportunities for significant allelic variation due to TE presence/absence polymorphisms among individuals. We used information on putative regulatory elements in combination with knowledge about TE polymorphisms in maize to identify TE insertions that interrupt existing accessible chromatin regions (ACRs) in B73 as well as examples of polymorphic TEs that contain ACRs among four inbred lines of maize including B73, Mo17, W22, and PH207. The TE insertions in three other assembled maize genomes (Mo17, W22, or PH207) that interrupt ACRs that are present in the B73 genome can trigger changes to the chromatin, suggesting the potential for both genetic and epigenetic influences of these insertions. Nearly 20% of the ACRs located over 2 kb from the nearest gene are located within an annotated TE. These are regions of unmethylated DNA that show evidence for functional importance similar to ACRs that are not present within TEs. Using a large panel of maize genotypes, we tested if there is an association between the presence of TE insertions that interrupt, or carry, an ACR and the expression of nearby genes. While most TE polymorphisms are not associated with expression for nearby genes, the TEs that carry ACRs exhibit enrichment for being associated with higher expression of nearby genes, suggesting that these TEs may contribute novel regulatory elements. These analyses highlight the potential for a subset of TEs to rewire transcriptional responses in eukaryotic genomes.

Keywords: transposable elements; cis-regulatory regions; ATAC-seq; DNA methylation; Zea mays

Introduction

Transposable elements (TEs) are highly repetitive DNA sequences found in most genomes. Variable genome size between related species has been partially attributed to the accumulation of TEs (Michael and Jackson 2013). The maize genome is replete with TEs, having >80% of ~2500 Mb of genomic space being composed of repetitive sequence and 64% annotated as complete TEs (Schnable *et al.* 2009; Jiao *et al.* 2016). TEs can be classified into two main orders based on their transposition intermediate, class I RNA retrotransposons that commonly proliferate through "copy and paste" transposition and class II DNA transposons that generally move through a "cut and paste" mechanism (Wicker *et al.* 2007). Barbara McClintock referred to these repetitive sequences as "controlling elements", encompassing their potential to impact and regulate genes (McClintock 1951). Transposition enables these TEs to move throughout the genome potentially influencing functional regions. TEs may insert into coding regions and cause direct influence on gene function and also may insert into existing regulatory regions or create new regulatory elements, resulting in altered gene expression (Lisch 2013; Chuong *et al.* 2017).

One mechanism of TE influence on gene expression is through the disruption of regulatory sequences. TEs in the maize genome are dispersed throughout the chromosome including gene-rich regions of chromosome arms (Baucom *et al.* 2009; Schnable *et al.* 2009). Due to this interspersion of genes and TEs, many TEs have the potential to influence the expression of genes. DNA transposons have been shown to display preferential insertion into genic regions (Dietrich *et al.* 2002; Liu *et al.* 2009; Vollbrecht *et al.* 2010; Springer *et al.* 2018) while retrotransposons appear to be more present in heterochromatic, gene-poor regions of the genome (Bennetzen 2000). In *Arabidopsis*, miniature inverted repeat

Received: September 29, 2020. **Accepted:** November 2, 2020

transposable elements (MITEs) often insert into the last exon of genes, which may cause more impact than ordinary intron insertions (Guo *et al.* 2017). A MITE DNA transposon, mPing in *Oryza sativa* was found to preferentially insert into the 5′ regions of genes (Naito *et al.* 2009). S elements in *Drosophila melanogaster* have been found to insert into the 5′ regions of several members of the Hsp70 heat shock gene family (Maside *et al.* 2002). MITEs and other TEs have been hypothesized to play an evolutionary role in altering gene expression through contributing regulatory elements (Wessler *et al.*, 1995; Bennetzen 2000; Lisch 2015).

TEs not only have the potential to disrupt regulatory sequence but can also introduce novel regulatory elements into new genomic locations (Feschotte 2008; Chuong *et al.* 2017). TE insertions can also result in changes in the location of regulatory elements relative to nearby genes (Zhao *et al.* 2018; Lu *et al.* 2019). It has been shown that TEs can impact gene expression through several examples in maize including *teosinte branched 1* (*tb1*), a gene responsible for the branching in the maize progenitor, teosinte (Studer *et al.* 2011). The regulatory region of *tb1* is within the intergenic space ∼60-kb upstream of the gene (Doebley *et al.* 1997; Clark *et al.* 2006; Briggs *et al.* 2007). An essential insertion of a retrotransposon Hopscotch acts as an enhancer of gene expression, resulting in the branching differences between maize and teosinte (Studer *et al.* 2011). Similar examples are observed in other species as well (Zhao *et al.* 2018; Nishihara *et al.* 2006; Lowe *et al.* 2007). The analysis of genes in the human genome has found evidence that TEs may contribute promoters (Jordan *et al.* 2003) or *cis*-regulatory regions (Sheffield *et al.* 2013). The existence of regulatory regions within TEs could represent examples of regulatory elements that have evolved to solely regulate the expression of the TE itself and examples in which the regulatory elements within the TE have been co-opted to regulate nearby genes (Chuong *et al.* 2017; Zhao *et al.* 2018).

The question of how TEs impact the genome has been considered from different perspectives since McClintock first discovered their existence. There are many examples in which detailed analyses of specific QTL have revealed the importance of TE insertions in creating altered gene expression (Zerjal *et al.* 2012; Zhang *et al.* 2012; Yang *et al.* 2013; Castelletti *et al.* 2014; Mao *et al.* 2015). There have been hints that certain families of TEs are associated with genes that exhibit stress-responsive expression (Makarevitch *et al.* 2015) and that many TEs exhibit dynamic, tissue-specific patterns of expression (Anderson *et al.* 2019b). There is evidence that a substantial number of accessible chromatin regions (ACRs) are found within TEs (Oka *et al.* 2017; Zhao *et al.* 2018; Lu *et al.* 2019) and in some cases these sequences can provide evidence for regulatory activity (Zhao *et al.* 2018).

To assess the mechanisms by which transposons might influence *cis*-regulatory elements, it is important to have an understanding of putative regulatory elements and transposon variation among genotypes. The availability of genome-wide identification of ACRs in B73 (Ricci *et al.* 2019) and high-quality information on shared and polymorphic TEs (Anderson *et al.* 2019a) provides new opportunities to address the potential impact of TEs on gene regulation in maize. We characterized hundreds of examples of B73 ACRs that are interrupted by a TE insertion in another genotype and thousands of examples of ACRs that are within annotated TEs. TE insertions into ACRs are associated with chromatin changes to the ACR in addition to the genetic change. Many of these ACRs within TEs show the evidence of functional enrichment. Through analyses of putative regulatory regions and TE polymorphisms, we can begin to evaluate how TEs may contribute to natural variation for gene expression in maize.

Methods

Annotation of genes and TEs

Whole-genome assemblies for B73 (Zm00001d) (Jiao *et al.* 2016), W22 (Zm00004b) (Springer *et al.* 2018), Mo17 (Zm00014a) (Sun *et al.* 2018), and PH207 (Zm00008a) (Hirsch *et al.* 2016) were used for genome-wide analyses. All analyses were done on assemblies of chromosomes 1–10 (the canonical maize chromosomes) while all un-placed scaffolds were disregarded due to the inability to compare these regions across genotypes. Filtered structural TE annotations (Stitzer *et al.* n.d.; Anderson *et al.* 2019a) were used.

Polymorphic TEs

Shared and non-shared TEs across genotypes were defined previously (Anderson *et al.* 2019a). Briefly, identification of shared and non-shared elements was determined through pairwise comparison between four maize inbred lines (B73, W22, PH207, and Mo17). Cross-genotype gene keys were generated using scrips available at https://github.com/SNAnderson/maizeTE_variation/gene-key_pipeline. Gene syntelogs were defined by a multi-approach method described in Anderson *et al.* (2019a,b) combining SynMap, Nucmer, and OrthoFinder. Search windows were defined by the closest, non-overlapping genes to the query TE with a syntelog in the genome being assessed. For comparison, 400-bp flanking tags were extracted for each annotated TE in the genome (for each genome assessed) centered at the start and end coordinates. These flank tags were mapped to the other genomes with the use of Burrows-Wheeler Aligner (BWA-MEM) (Li and Durbin 2009) in paired-end mode. Further characterization was performed on those elements with tags mapped completely within the search window. Non-shared site-defined TEs were defined by the alignment of only the outer 200 bp of the flank tags where the distance between tags was less than twice the TSD length for the superfamily. This resulted in a total of 69,292 non-shared site-defined elements across all pairwise comparisons used for analyses (Anderson *et al.* 2019a).

A total of 509,629 non-redundant TEs defined in at least one of the B73, Mo17, PH207, or W22 structural TE annotations were assigned as present or absent in 509 of the WiDiv inbred genotypes (Hansey *et al.* 2011; O'Connor *et al.* 2020). Methods for classification of present/absence TEs are described in O'Connor *et al.* (2020) (BioRxiv 10.1101/2020.09.25.314401). Briefly, two points of reference, 10 bp over left and right inner edges of a TE, were used to determine TE status in a particular genotype. TEs with coverage ≥8 across both inner edges were classified as present while TEs with coverage <7 across both inner edges were classified as absent. All other TEs were classified as ambiguous. All TEs defined as present and absent in at least one other genotype were maintained for downstream analyses (Presence/absence variation (PAV) calls across the 509 inbred lines for each TE can be found in the DRUM database: http://hdl.handle.net/11299/216935). Data presented in O'Connor *et al.* (2020) only use a subset of this TE list based on a frequency threshold of genotypes with an ambiguous classification. Sequencing data (with >20× coverage) for each of the 509 inbred maize genotypes are available at SRA (BioProject PRJNA661271).

Methylation data

In this study, we utilized previously generated WGBS data for B73 seedling shoot, PH207 seedling shoot, Mo17 seedling leaf, and W22 seedling leaf. Trim_glore (Martin 2011) was used to trim adapter sequences, and read quality was assessed with the default parameters and paired-end reads mode. Reads that passed

quality control were aligned to the B73v4 genome (non-B73 genotypes were also aligned to their corresponding genome assemblies). Alignments were conducted using BSMAP-2.90 (Xi and Li 2009), allowing up to 5 mismatches and a quality threshold of 20 (-v 5 -q 20). Duplicate reads were detected and removed using picard-tools-1.102 (Picard Tools – By Broad Institute, 2018) and SAMtools (Li *et al.* 2009). Conversion rate was determined using the reads mapped to the unmethylated chloroplast genome. The resulting alignment file, merged for all samples with the same tissue and genotype, was then used to determine the methylation level for each cytosine using BSMAP tools. Methylation ratios for 100-bp non-overlapping sliding windows across the B73v4 genome in all three sequence contexts (CG, CHG, and CHH) were calculated (#C/(#C + #T)). Each 100-bp window was categorized as methylated (≥40%), intermediate (20–40%), or unmethylated (≤20%) based on the CHG methylation level.

ATAC-seq data

In this study, we utilized previously generated seedling shoot ATAC-seq data for B73 (Ricci *et al.* 2019). Raw reads were trimmed with Trimmomatic v0.33. Reads were trimmed for NexteraPE with a maximum of two seed mismatches, a palindrome clip threshold of 30, and a simple clip threshold of 10. Reads shorter than 30 bp were discarded. Trimmed reads were aligned to the *Zea mays* AGPv4 reference genome 44 using Bowtie v1.1.147 with the following parameters: "bowtie -X 1000 -m 1 -v 2 –best –strata". Aligned reads were sorted using SAMtools v1.3.1, and clonal duplicates were removed using Picard version v2.16.0 (http://broadinstitute.github.io/picard/).

Identification of ACRs

MACS2 was used to define ACRs with the "–keep-dup all" function and with ATAC-seq input samples (Tn5 transposition into naked gDNA) as a control. The ACRs identified by MACS2 were further filtered using the following steps: (1) peaks were split into 50-bp windows with 25-bp steps; (2) to quantify the accessibility of each window, the Tn5 integration frequency in each window was calculated and normalized with the average integration frequency across the whole genome to generate an enrichment fold value; (3) windows with enrichment fold values passing a cutoff (25-fold) were merged together by allowing 150-bp gaps; and (4) to remove possible false positive regions, small regions with only one window were filtered for lengths >50 bp. The sites within ACRs with the highest Tn5 integration frequencies were defined as ACR "summits".

For the functional analysis of single nucleotide polymorphism (SNP), HiChIP, STARR-seq, and eQTL data, we utilized the same methods as described in Ricci *et al.* (2019). The difference lies in the subset of data that was used to focus on TE-ACRs *vs* non-TE-ACRs opposed to all distal ACRs in the genome.

Determination of TE-ACR overlap

TE-ACRs were defined by an overlap of B73 ACR coordinates with the structural TE annotation coordinates. Each ACR was assigned to a single TE using bedtools closest based on the disjoined TE coordinates file. For those with a partial overlap of multiple TEs, the ACR was assigned to the TE with the greatest overlap. Complete overlaps were defined by >80% of the ACR length overlapping a TE.

Identifying TE insertions into ACRs

Site-defined TE polymorphisms with the TE present in Mo17, W22, and/or PH207 and absent in B73 were utilized to identify TE insertions into ACRs. Bedtools intersect was run with all defined B73 ACRs and the site-defined insertions, using the B73 insertion site coordinates. Any site-defined TE in Mo17, PH207, and/or W22 that had an insertion site within the coordinate range of a B73 ACR was characterized as a TE insertion into an ACR for further analyses.

A set of control regions was generated as a genome-wide proxy for potential accessible regions. The genome was a subset to "mappable" sequence determined by WGBS read coverage and used as the input to bedtools shuffle along with the identified ACRs. Output contains the same number of regions with the same lengths as the ACR input file randomly placed across the mappable genome. These regions are used as a control for the frequency of insertions into accessible regions.

Analysis of methylation at TE insertion sites

Methylation for each TE insertion was defined for the TE-present genotype (Mo17, PH207, or W22) and the TE-absent genotype (B73). Changes in methylation were identified by comparing 100-bp bin CG methylation of the ACR in B73 to CG methylation levels flanking the insertion site in the genotype present for the TE. The position of the insertion was determined by its location in the ACR by quartiles with the 1st and 4th quartiles being insertions at the edge of the ACR and the 2nd and 3rd quartiles defined as insertions into the middle of the ACR.

Analysis of methylation at ACRs across genotypes

Gene anchor files have been one-to-one gene syntelogs pairwise between B73, Mo17, PH207, and W22. Gene key files are available at https://github.com/SNAnderson/maizeTE_variation and were filtered to only one-to-one gene matches. Bedtools closest upstream and downstream, ignoring overlaps, was run for each B73 ACR relative to gene anchor files between B73 and PH207, W22, and Mo17. The search window was defined by the closest upstream and downstream non-overlapping genes in the query genome on either side of the ACR sequence that has a unique syntelog in the target genome. BLAST was run for each B73 ACR sequence to PH207, W22, and Mo17 to identify sequence similarity in the search window for the corresponding genotype. The sequence coordinates were identified and bedtools overlap was run against the 100-bp WGBS data for that genotype. The methylation state of the B73 ACR was compared to the methylation levels of the matching sequence in PH207, W22, and Mo17 (based on WGBS data aligned to the corresponding genome assembly). The ACR was characterized as methylated if the average level of CHG methylation was >40% and unmethylated if the average level of CHG methylation was <20%. A change in methylated was identified by an ACR characterized as unmethylated in B73 having a methylated state in another genotype.

Gene expression analyses

RNA-seq datasets (Hirsch *et al.* 2014; Kremling *et al.* 2018) were used to assess expression levels across 284 genotypes and 8 tissues. To assess gene expression variation, the closest gene to each TE was determined in B73 and the expression of that gene was associated with the presence or absence of the TE in each of the 284 genotypes. Each element containing an ACR or inserting into an ACR was assigned to the closest B73v4 annotated gene (in either direction) using bedtools closest. Only one assignment was given for each TE and any TE annotated as containing the full sequence of a gene was removed from the analysis. For those with distal ACRs, HiChIP data were used to assign the gene if an interaction was identified (Supplementary Table S2/S3). TE presence

impact was determined for each TE–gene pair by averaging the expression values for TE-present genotypes and TE-absent genotypes and the log2(present/absent) value was calculated. To account for biases in the number of genotypes with each TE as present or absent, a *t*-test was performed to determine the *P*-value for each gene in each tissue.

Data availability

In this study, we utilize datasets that are available through the following accessions: SRX4727413, SRR8738272, SRR8740852, and BioProject PRJNA661271. The TE polymorphism data used for analysis in this work are available at http://hdl.handle.net/11299/216935.

Supplementary material is available at figshare DOI: https://doi.org/10.25386/genetics.13182986.

Results

To assess potential impacts of TEs on putative regulatory regions in the maize genome, we used a set of 32,421 previously characterized maize ACRs identified using an Assay for Transposase-Accessible Chromatin with sequencing, hereafter referred to as ATAC-seq (Ricci *et al.* 2019). Roughly similar numbers of ACRs were found within genes (12,587), proximal regions (within 2 kb of genes—9183), and distal regions (>2 kb from nearest gene—10,651). Ricci *et al.* (2019) documented evidence to support the functional relevance of distal ACRs through the enrichment of genetic variants underlying morphological and expression variation (eQTL and GWAS), chromatin–chromatin (HiChIP) interactions, and self-transcribing active regulatory region sequencing (STARR-seq) enhancer activity. We sought to investigate the role that TEs might play in regulating gene expression by disrupting ACRs within the maize genome or in carrying ACRs within TEs (Figure 1A/B). To monitor TE insertions within TE-ACRs, we focused on the set of ACRs identified within the B73 genome (Ricci *et al.* 2019) and documented the TE insertions in these regions within the W22, Mo17, or PH207 genomes (Figure 1C). The TEs that contain an ACR (>80% of ACR within the TE) were determined by comparing the coordinates of ACRs within the B73 genome with the B73 TE annotations (Figure 1D). The set of TE insertions into ACRs and TEs containing ACRs were further characterized to understand how these changes might influence chromatin states and regulation of nearby genes.

Identification of TE insertions into ACRs

Of the 348 non-redundant instances of TE insertions into B73-defined ACRs, 176 TE insertions were found in Mo17, 82 insertions in PH207, and 158 insertions in W22. To determine the number of TE insertions expected by chance, we used a random set of genomic regions with similar size distribution as the ACRs. We observe significantly (Fisher's exact *P*-value = 4.286e−07) more TE insertions in ACRs compared to the random regions (Supplementary Figure S1A). The TEs that inserted were primarily terminal inverted repeat (TIR) DNA transposons with fewer examples of long terminal repeat (LTR) retroelements and Helitrons (Figure 1 and Supplementary Figure S1B). Several TIR elements have been found to be enriched for insertions within accessible chromatin (Kolkman *et al.* 2005; Liu *et al.* 2009; Han *et al.* 2013; Noshay *et al.* 2019). The insertions into ACRs are highly enriched for members of the hAT (DTA) and Mutator (DTM) DNA transposon superfamilies (Supplementary Table S1) of TIR elements (Supplementary Figure S1C). The TE insertions located within ACRs tended to represent relatively young TEs based on LTR similarity (Supplementary Figure S1D).

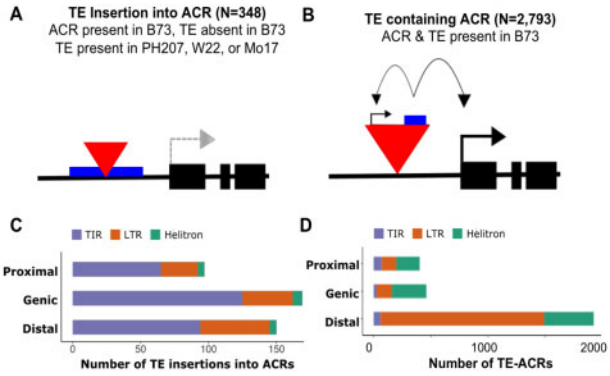

Figure 1 An overlap of TEs and ACRs. Schematic representation of the identified ACRs (blue) in the B73 maize inbred line and their interaction with TEs (red) and the potential impact on nearby genes. (A) B73 ACRs that have a site-defined TE insertion in Ph207, Mo17 or W22. (B) B73 ACRs that are found within B73 TE sequence. (C) The number of TE insertions (as shown in A) in PH207, Mo17, or W22 into each ACR category (characterized by their position relative to annotated genes as genic, proximal, or distal) of ACR based on site-defined insertion sites in B73. Colors represent TE order. (D) Number of TE-ACRs (as shown in B) by location relative to genes and TE order.

TE insertions into ACRs can result in altered chromatin

The ACRs represent regions of accessible chromatin and also lack DNA methylation (Ricci *et al.* 2019). The insertion of a TE in another haplotype could result in not only a genetic change to the DNA sequence but also to changes in chromatin modifications or accessibility. DNA methylation data were generated for the same tissue type used for ATAC-seq in both B73 and PH207. There are 82 examples of PH207 TE insertions within B73 ACR regions and these were used to investigate the frequency of DNA methylation presence within the region classified as an ACR in B73. Specifically, we assessed the frequency of DNA methylation gains on one (uni-directional) or both (bi-directional) sides of the TE insertion (Figure 2A). In many cases, the insertion of a TE within an ACR is not associated with increased methylation of the regions with homology to the B73 ACR (Figure 2B). However, for 37% of the TE insertions within ACRs, there are DNA methylation gains in the haplotype with the TE insertion (Figure 2C). TE insertions that are located within the outer quartiles of the ACR often exhibit methylation gains only on one side of the TE, typically in the region closer to the edge of the ACR (Figure 2D). These analyses were solely focused on TE insertions within the B73-defined boundaries of the ACR. An analysis of 257 additional TE insertions (present in PH207, Mo17, or W22) located within 200 bp of the ACR (present in B73) identified 30 additional examples in which a TE insertion near an ACR was associated with DNA methylation gains within the ACR. Together these analyses suggest that a subset of the TE insertions within, or near, ACRs are associated with changes to the DNA methylation state of the region and are likely associated with changes in chromatin accessibility.

Identification of ACRs within TEs

In addition to the potential for TEs to disrupt existing ACRs, they have the potential to carry sequences that lead to an accessible chromatin state and potentially move these sequences to new genomic locations (Figure 1B). We focused on characterizing

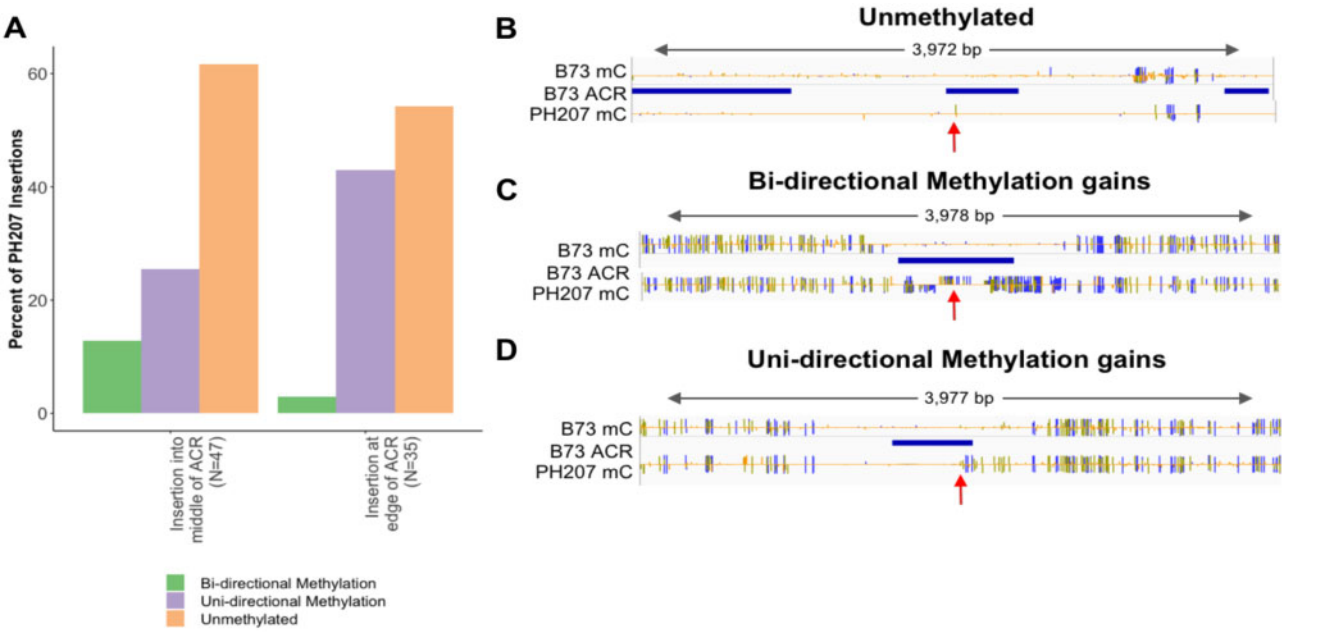

Figure 2 Methylation changes due to TE insertions in PH207. (A) For every PH207 site-defined TE insertion into a B73 ACR, the PH207 methylation status is defined as unmethylated (region remains unmethylated just as it was in B73), uni-directional methylation (methylation gain on one side of the insertion site), or bi-directional methylation (methylation gain on both sides of the insertion site). Insertions are broken into those that insert into the middle of an ACR (quartile 2 or 3) or those that insert into the edge of an ACR (quartile 1 or 4). WGBS data for B73 and PH207 were aligned to the B73 genome to visualize. IGV views display methylation level tracks (blue is CG, green is CHG, and yellow is CHH), ACR region tracks, and TE insertion sites indicated by red arrows. These are shown for each methylation status: (B) unmethylated, (C) bi-directional methylation, and (D) uni-directional methylation.

examples of the ACRs that are identified in the B73 genome located within or overlapping annotated TEs. Of the 32,421 identified ACRs in maize, 4590 have at least a partial overlap with an annotated TE (Table 2). It is worth noting that this is likely an underestimate of the number of true ACRs within TEs as the identification of ACRs relied upon uniquely mapping reads (Ricci *et al.* 2019). Many TEs are repetitive and have enough similarity to other family members to preclude uniquely mapping reads, which means that the number detected using unique mapping represents only a subset of actual accessible regions within TEs (Supplementary Figure S5). In both leaf and ear tissue, there is no evidence for enrichment of unique mapping reads in ATAC-seq data suggesting the presence of accessible chromatin within repetitive regions (Supplementary Figure S5A). On a per-TE family basis, in which we could determine the number of reads that map to a family (both multiple mapping and unique mapping reads), there is evidence for some families with substantially more multi-mapping reads (Supplementary Figure S5B). However, the multi-mapping reads cannot be attributed to a single genomic location and, therefore, we focused on the ACRs classified based on unique mapping reads for the remainder of our analyses.

Among the 4590 TE-ACRs, there are 2793 examples in which the majority (>80%) of the ACR is located within the TE and another 1797 that have partial overlap (<80%) (Table 2 and Supplementary Figure S3A). These 1797 partial overlaps may represent instances in which the ACR within the TE includes some adjacent sequence or may represent instances in which the TE inserted into an existing ACR and the accessible region spreads to encompass a portion of the TE. ACRs within TEs are more common for distal ACRs than for the other types of ACRs, especially for ACRs with majority (>80%) overlap with a TE (Supplementary Figure S3A). The partial overlaps of ACRs with TEs have a high frequency of TIR elements, while the majority (>80%) overlap TE-ACRs have much higher frequencies of LTR elements (Supplementary Figure S3A). Given the potential for the partial overlaps to represent instances of TE insertion into or near ACRs, rather than carrying the ACR within the TE, we focused on the majority (>80%) overlaps for the analyses of ACRs within TEs.

Table 1 B73 ACRs majority overlapping (>80%) or partially overlapping (<80%) annotated TEs

	Genic	Proximal	Distal
Total	12,587	9183	10,651
LTR	138 (93)	130 (94)	1428 (225)
TIR	25 (382)	72 (387)	63 (376)
Helitron	301 (90)	203 (74)	433 (76)
Total TE	464 (565)	405 (555)	1924 (677)

Values in brackets represent partial overlaps (<80%).

The 2793 examples of majority TE-ACR overlap mostly (69%) comprise examples of distal ACRs (Figure 1D). Even though only 0.98% of all maize TEs contain an ACR, 19% of the distal ACRs are located within a TE (Table 2). Given an expectation that TEs would not contain accessible chromatin, this represents a large number of unexpected ACRs within TEs. However, if we assume that ACRs are randomly located in genomic sequence, then the fact that 19% of distal ACRs are found within TEs is actually substantially fewer than expected (72% of random distal regions with size distribution similar to ACRs overlap a TE) given the amount of sequence attributed to TEs in the maize genome. The distal ACRs were further classified based on the patterns of several chromatin modifications into four groups; K-acetyl enriched, H3K27me3 enriched, transcribed, and unmodified (Supplementary Figure S3B) (Ricci *et al.* 2019). The TEs containing ACRs are enriched (chi-square *P*-value <2.2e−16) for the transcribed class, which is characterized by H3K4me3 and

Table 2 RNA-seq and TE PAV dataset summaries

Dataset	Number of tissues	Genotypes with TE calls	TE insertion ($N = 377$)		TE-ACR ($N = 2182$)	
			Significant (+/−)	Outliers (+/−)	Significant (+/−)	Outliers (+/−)
Kremling *et al.* (2018)	GRoot	91	2/1	16/17	51/2	214/59
	GShoot	91	3/10	0/27	55/4	204/54
	Kern	84	4/3	15/23	67/2	240/60
	L3Base	87	2/4	19/25	54/4	197/65
	L3Tip	86	5/1	19/22	44/6	281/60
	LMAD	54	3/0	17/27	30/8	265/86
	LMAN	94	0/3	14/32	52/11	256/73
Hirsch *et al.* (2014)	Seedling	230	½	2/7	57/14	105/22
Non-redundant sum	All of the above	259	9/12	57/86	153/37	667/295

H3K36me3 along with acetylation marks and low DNA methylation levels similar to patterns seen in the promoters of expressed genes. This suggests that at least a portion of the ACRs found within TEs may represent promoters for expressed TE products. Prior work monitored the expression of TEs in a variety of B73 tissues, including pollen and other reproductive tissues (Anderson *et al.* 2019b). Of the TEs containing an ACR classified as transcribed, 48% show observable expression levels in at least one tissue (Supplementary Figure S3C). The TEs containing ACRs in the other classes (chromatin marked and unmodified) have lower frequencies of expressed elements but are still expressed more often than non-ACR TEs (Supplementary Figure S3C).

We investigated the potential that TE-ACRs would be found primarily near highly expressed genes. Using expression data from the same tissue used to perform chromatin accessibility profiling, the genes were divided into not expressed ($n = 13,956$) and four expression quartiles ($n = 6,262$ in each quartile) (Supplementary Figure S4). As expected, expressed genes were enriched for the presence of ACRs within 5 kb of the TSS and highly expressed genes were more likely to have an ACR than low-expressed genes (Supplementary Figure S4A). However, only a small proportion of genes in any group had a TE-ACR within 5kb of the TSS. Highly expressed genes were slightly more likely to have a TE-ACR nearby, but in general expressed genes have similar overall numbers of TEs with and without ACRs (Supplementary Figure S4B). This suggests that some of the TE-ACRs may occur due to proximity to highly expressed genes and also reveals that similar numbers of silent or lowly expressed genes also contain TE-ACRs.

Evidence for potential functional regulatory elements within TEs

Ricci *et al.* (2019) used several approaches to provide evidence for functional impacts of distal ACRs. Focusing on the 10,651 distal (>2 kb from nearest gene) ACRs, we sought to determine whether there were differences in the support of functional impact for ACRs within TEs (TE-ACR) compared to ACRs located outside of TEs (non-TE-ACR). The frequency of SNPs is reduced within ACRs, and this effect becomes even more pronounced when focusing on the TE-ACRs (Figure 3A). The analysis of the frequency of GWAS-associated SNPs revealed enrichment within both TE-ACRs and non-TE-ACRs (Figure 3B). TE-ACRs also show enrichment for eQTL, although the level of enrichment is not as strong as observed for non-TE-ACRs (Figure 3C). The difference in the level of eQTL enrichment for TE-ACRs and non-TE-ACRs could be due to the differences in composition among the four chromatin classes of ACRs. The transcribed ACRs generally have lower enrichment than observed for some of the other classes (Supplementary Figure S6). For ACRs to influence expression, they would likely need to interact with nearby gene promoters. HiChIP analysis of chromatin interactions reveals similar enrichment for ACR-genic interactions for both TE and non-TE-ACRs (Figure 3, D and E). STARR-seq can identify sequences that can provide functional enhancer activity. STARR-seq analysis of maize accessible chromatin fragment activities in maize leaf protoplasts showed similar levels of enrichment for enhancer activity for TE and non-TE-ACR sequences (Figure 3F).

Enrichment for certain TE families containing ACRs

TEs are classified into order, superfamily, and family based on transposition mechanism, structural components, and sequence similarity. The ACRs that are located within TEs may represent TE family-specific properties in which multiple members of the same family contain an ACR or could represent instances in which the local chromatin neighborhood for a specific TE insertion allows the formation of an ACR. There are 356 (12.7%) of the 2793 TE-ACRs that are located within single-member TE families, which is much greater than the overall frequency (1.5%) of single copy TEs in the genome. Among the remaining 2437 TE-ACRs that are within multi-member TE families, 557 are only in one of the TEs in the family containing an ACR. This suggests that the majority of TE-ACRs are not a reproducible feature of the family members. A caveat to these results is the repetitive sequences, which would not have been captured through the unique mapping ATAC-seq analysis and therefore additional members of a family may contain ACRs (Supplementary Figure S5B).

There are examples of TE-ACRs that are found in multiple members of a TE family. There are 112 TE families with at least two members with an ACR. There are only 10 of these families (with at least 3 elements) in which >30% of the elements have an ACR (Supplementary Figure S7A). These examples of TE families with multiple members with ACRs were identified based on the utilization of unique mapping reads. It is quite possible that additional members of these families may contain ACRs that were not identified because they are in regions that are highly similar in multiple TEs and therefore are multi-mapping. Two families in particular, RLX00813 and RLX01441, were found to display increased coverage when multi-mapping was allowed (Supplementary Figure S7B).

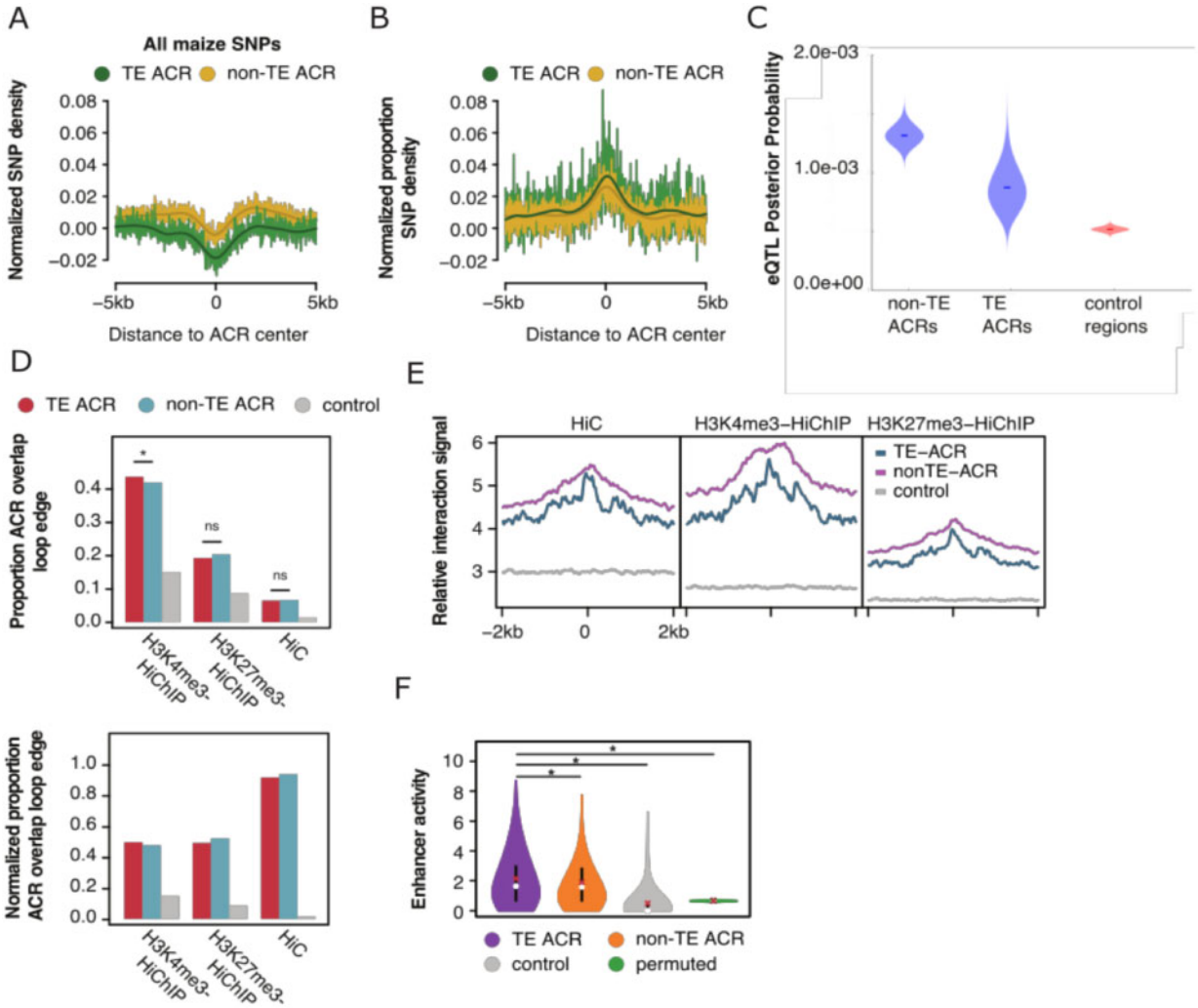

Figure 3 Functional differences between TE and non-TE-ACRs among distal ACRs. (A) Normalized (control) SNP density among maize inbred lines averaged across 10-kb regions centered on TE and non-TE-ACRs. (B) Proportion of GWAS hits (out of all maize SNPs) normalized by control enriched within 10-kb windows centered on TE and non-TE dACRs. (C) eQTL posterior probability for TE and non-TE-ACRs compared to control regions. (D) Contrasts between the proportions of dACRs overlapping an I-G loop between TE-ACRs and non-TE-ACRs. Chi-square, *P-value <0.05. (E) Relative enrichment of chromatin interaction tags across 4-kb windows centered on TE-ACRs and non-TE-ACRs across the three types of chromatin loops. (F) Distribution of enhancer activities for dACRs split by the presence/absence of TEs, control regions ($n = 4406$), and the means of a permutation (10,000×). Statistical differences between TE and non-TE-ACRs were evaluated with Mann–Whitney rank sum test. Statistical differences between distribution means and permuted regions were estimated as empirical *P*-values. ns, not significant; $^*P < 0.05$.

ACRs within TEs show variable DNA methylation patterns among genotypes

In general, TEs are considered to have quite high levels of DNA methylation, but ACRs typically lack DNA methylation (Oka *et al.* 2017; Lu *et al.* 2019; Ricci *et al.* 2019). The presence of ACRs within TEs led us to investigate the DNA methylation level of these sequences. We found that while TEs containing an ACR show quite high levels of DNA methylation throughout most of the TE, the ACR section is essentially unmethylated (Figure 4, A and B). Visual inspection of several examples reveals that the ACR region represents a small window of unmethylated DNA within the largely methylated TE (Figure 4, C and D).

We hypothesized that the presence of an unmethylated region within a TE might be somewhat unstable and could be subject to changes in the DNA methylation state among different haplotypes at a higher frequency than ACRs not located within TEs. An analysis was performed using a set of B73 ACRs that have a matching sequence at a syntenic location in PH207, Mo17, or W22 and have DNA methylation data available for both genotypes. These include ACRs within TEs that are present in both genomes and ACRs that are present in non-TE sequence (non-TE-ACRs). While <3% of the non-TE-ACRs exhibit gains of CG methylation across each of the genotypes, there are over 12% of the ACRs that are located within TEs that exhibit high levels of CG methylation (Figure 5A). Visual inspection of several loci suggests gains of both CG and CHG methylation over the full ACR sequence in these examples (Figure 5, B and C). These observations suggest that ACRs within TEs may exhibit less stability among genotypes than ACRs in non-TE regions of genomes.

TE presence association with gene expression

Polymorphic TEs that interrupt an ACR or create novel ACRs in some haplotypes have the potential to influence the expression of nearby genes. To assess the potential for these polymorphic TE-ACR interactions to influence gene expression, we sought to associate the presence/absence of TEs with the changes in

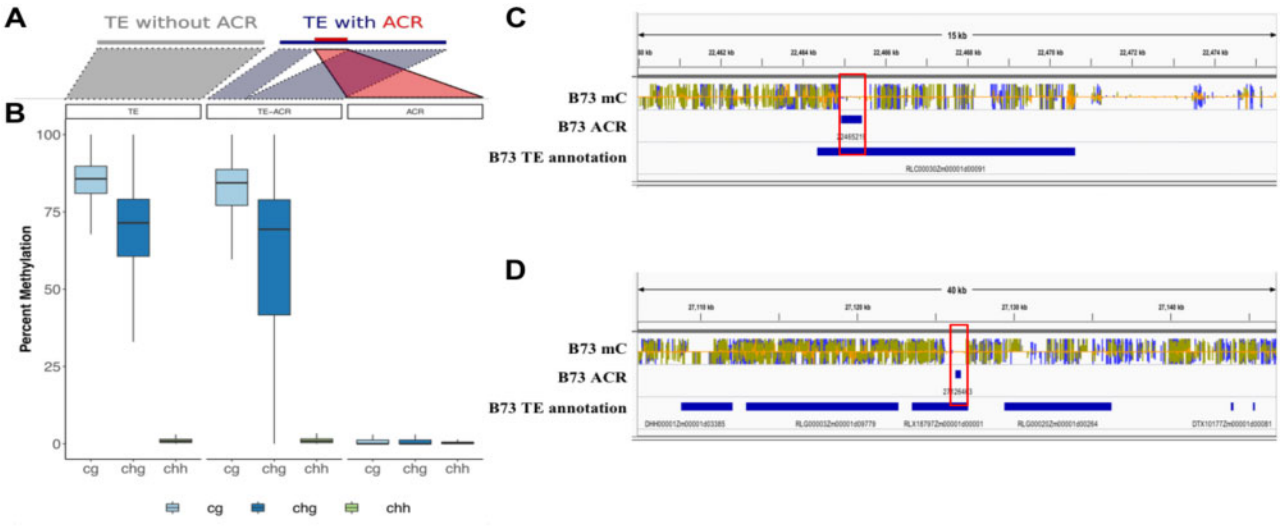

Figure 4 TE-ACR methylation patterns. (A) Schematic representation of a TE without an ACR (gray) and a TE containing an ACR (blue) with the ACR sequence shown in red. (B) Methylation levels of TEs without ACRs, TEs with an ACR (excluding ACR bins), and ACRs showing the trend that TEs maintain similar levels of high CG and CHG methylation with and without an ACR but the ~300-bp region of an ACR is unmethylated. (C/D) IGV view of TE with an ACR and the methylation levels (CG, blue; CHG, green; CHH, yellow) over a majority of the TE and absence over the ACR.

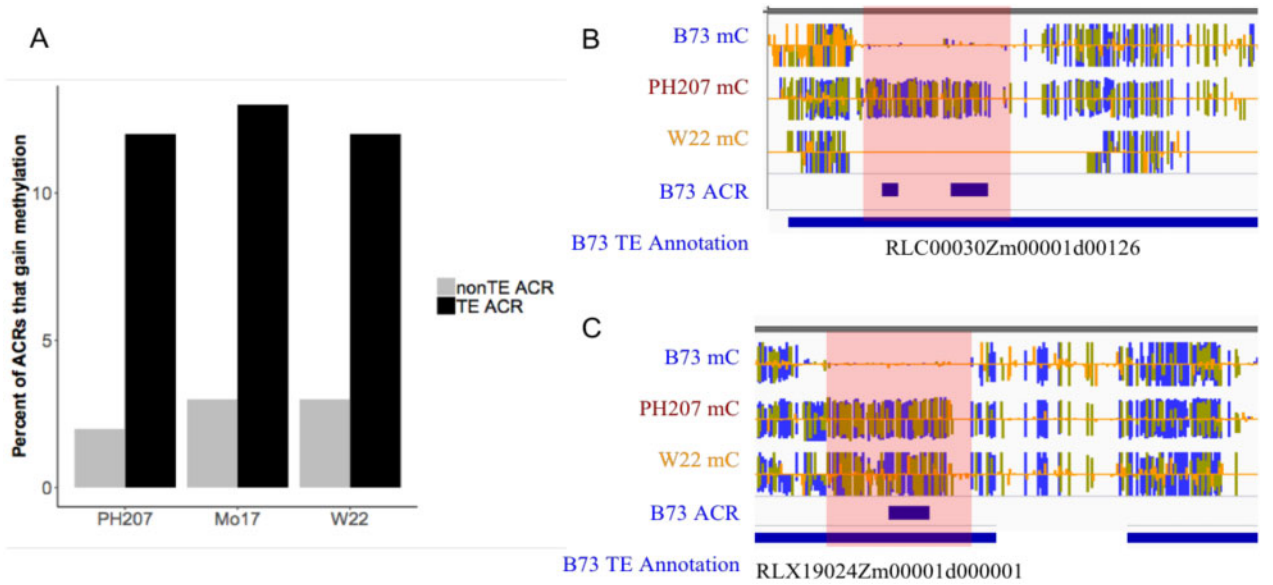

Figure 5 Unmethylated (open chromatin) regions in TEs are less stable than non-TE open chromatin regions. (A) Percent of ACRs that gain methylation in PH207, Mo17, or W22 for non-TE-ACRs (gray) and TE-ACRs (black). (B/C) IGV view of B73 TE annotation with unmethylated ACR in B73 and the same region as methylated in PH207 and/or W22. Methylation tracks show CG methylation in blue, CHG methylation in green, and CHH methylation in yellow.

relative expression levels for nearby genes in panels of diverse germplasm. De novo assembled genome sequences of B73, Mo17, PH207, and W22 were used to generate de novo TE annotations in these four genomes (Anderson *et al.* 2019a). The presence or absence of these TEs was assessed in a larger (>500 inbreds) panel of diverse maize lines using alignments of whole-genome shotgun sequencing reads to the TE-flanking sequence junctions (O'Connor *et al.* 2020). This approach provides robust assignments of presence or absence for many genotypes, but in some cases, there is no clear evidence and the TE status is classified as ambiguous in that genotype. The TE polymorphism information was used to investigate variation in gene expression in several RNA-seq datasets (Hirsch *et al.* 2014; Kremling *et al.* 2018; Mazaheri *et al.* 2019). Each of these datasets included samples from a panel of genotypes that were collected at similar tissue stages.

Each polymorphic TE that disrupts a B73 ACR or contains an ACR in B73 was assigned based on HiChIP interactions or proximity to the nearest gene. TE–gene pairs where the gene is present completely within an annotated TE were disregarded for this analysis. We then assessed the difference in expression for genotypes with or without the TE insertion across the two datasets incorporating 284 genotypes and 8 tissues (Table 1 and Supplementary Figure S9), allowing separate tests of potential associations between TE polymorphisms and expression level in multiple tissues. We initially focused on the set of 377 TE insertions into an ACR, which we hypothesized may result in reduced expression for the nearby gene. The majority of these TE

insertions into ACRs have limited associations with the expression of nearby genes. There are 21 instances (5.6% of all TE–gene pairs) in which we found a significant (q-value <0.05 and >2-fold change) change in expression for the nearby gene (Table 1). These include 9 genes in which higher expression was observed for the haplotype containing the TE insertion, and 12 examples of lower expression when the TE is present. In 10 of the 21 significant associations, we found a significant association between the presence of the TE and expression levels in multiple tissues. In addition to the genes with significant associations, we noticed that there is an apparent excess of many "outlier" expression states for which the genotype with (or without) the TE has a >30-fold change in expression, but there is limited statistical significance because one of the haplotypes is rare (Supplementary Figure S9A). To determine if there is a significant excess of these outliers, we performed separate permutation tests in which the genotype-expression or genotype-TE presence classifications were randomized. These were separately performed for each of the expression datasets and were used to determine the number of significant or outlier expression changes expected by chance within this data structure (Figure 6A). The TE insertions into ACRs consistently exhibit more outliers than expected by chance with reduced expression of the haplotype with the TE present for each of the expression datasets (Figure 6A).

We next assessed the 2182 polymorphic insertions of TEs containing ACRs near genes, which were hypothesized to have positive influences on the expression of the nearby gene. There were 190 significant associations (8.7% of all tested TE–gene pairs) and 81% of these significant associations exhibit higher expression for the nearby gene (Supplementary Figure S9B and Table 1). Many (49%) of the significant positive associations between the presence of the TE and the expression of the nearby gene were identified in multiple tissues while fewer (18%) of the negative associations were identified in multiple tissues. Figure 6, C and D, shows two examples of a TE located near a maize gene with significant positive associations with expression in multiple tissues. In both of these examples, there are HiChIP interactions between the ACR within the TE and the nearby gene based on data from Ricci *et al.* (2019). The permutation tests identify very few significant associations (Figure 6B). The analysis of rare outlier expression states also reveals an excess of positive associations in which the haplotype containing the TE exhibits a higher expression level (Figure 6B). To further support the *cis*-regulatory variation observed at the examples of significant associations between the presence of TE-ACRs and the expression of nearby genes, we evaluated allelic bias for expression in F1 hybrids. Prior work had generated allele-specific expression for 23 tissues in the B73 × Mo17 F1 hybrid (Zhou *et al.*, 2019). There are 26 polymorphic TE-ACR insertions in B73-Mo17 with significant associations with expression and allele-specific data available. When we investigate tissue types most closely related to the tissue with significant associations, we find significant allelic expression bias for 19 of these 26 genes in the predicted direction (Supplementary Figure S10). Most of the seven genes without significant allelic bias still exhibit a bias in the expected direction but did not contain sufficient sequencing depth to provide evidence for significant effects. This further confirms the presence of *cis*-regulatory variation for these loci.

Discussion

Many eukaryotic genomes show evidence for both recent amplification of TEs and turnover of elements through deletions (Bennetzen and Kellogg 1997). Insertions of transposons into genes or regulatory elements can lead to loss-of-function mutations, which are presumed to be primarily deleterious. However, there is growing evidence that TEs may also contribute to the rewiring of transcription of nearby genes (Weil and Martienssen 2008; Feschotte 2008; Lisch 2013; Chuong *et al.* 2017). Transposon insertions that affect expression of a nearby gene are the molecular basis for allelic variation at several loci important for maize domestication and improvement (Studer *et al.* 2011; Yang *et al.* 2013; Castelletti *et al.* 2014). There are also examples in maize and other species in which transposon insertions may influence regulatory influences on nearby genes (Jiang *et al.* 2004; Cavrak *et al.* 2014; Makarevitch *et al.* 2015; Zhao *et al.* 2018). While specific examples have been identified, the genome-wide frequency for these TE influences has not been characterized. Advances in our knowledge of genome-wide TE polymorphisms (Stitzer *et al.* 2019; Anderson *et al.* 2019a) and the identification of proximal and distal putative *cis*-regulatory elements (Oka *et al.* 2017; Zhao *et al.* 2018; Ricci *et al.* 2019) provided an opportunity to assess the mechanisms and frequency by which TEs may create regulatory variation

In this study, we focused on two potential ways in which TEs might influence the expression of nearby genes: the disruption of regulatory regions and the introduction of novel sequences that may act as regulatory sequences. Insertions into regions of accessible chromatin might be expected to often result in reduced expression of nearby genes or altered patterns of expression. In contrast, TEs that contain ACRs may be mobile enhancers that affect the expression of both the TE promoter and nearby gene promoters. Several studies have found that putative enhancers can be found within TEs in the maize genome (Oka *et al.* 2017; Zhao *et al.* 2018). We were interested in assessing how frequently the polymorphic insertions could be associated with variable expression for nearby genes to understand the potential for TE polymorphism to generate regulatory diversity. It is worth highlighting the fact that truly assessing the potential for TEs to influence regulation in natural populations may be complicated by the potential fitness consequences of polymorphic TE insertions. If a TE insertion results in significant deleterious or beneficial consequences, the allele will likely be a target of selection. Recent studies have found that there are likely many examples of rare deleterious expression states in domesticated maize populations (Kremling *et al.* 2018) and therefore we monitored both common and rare expression states associated with TE polymorphisms.

Potential for TEs to reshape chromatin and the epigenome

Active transposition of TEs results in genetic changes including disruption of genes or regulatory elements as well as potential genomic instability due to chromosome breaks or illegitimate recombination. To limit these deleterious events, most genomes have evolved mechanisms to restrict active transposition, including epigenetic silencing through chromatin modifications such as DNA methylation (Hollister and Gaut 2009; Lisch 2013; Springer *et al.* 2016). This results in highly methylated TEs in plant genomes (Niederhuth *et al.* 2016) and has been observed to spread outside of the TE sequence to surrounding DNA sequences in some cases (Eichten *et al.* 2012; Choi and Purugganan 2018; Noshay *et al.* 2019; Wyler *et al.* 2020). As TEs insert into putative regulatory regions, the question becomes not only how the presence of new DNA sequence impacts this region but also the potential for alteration in chromatin patterns. The TE insertion into

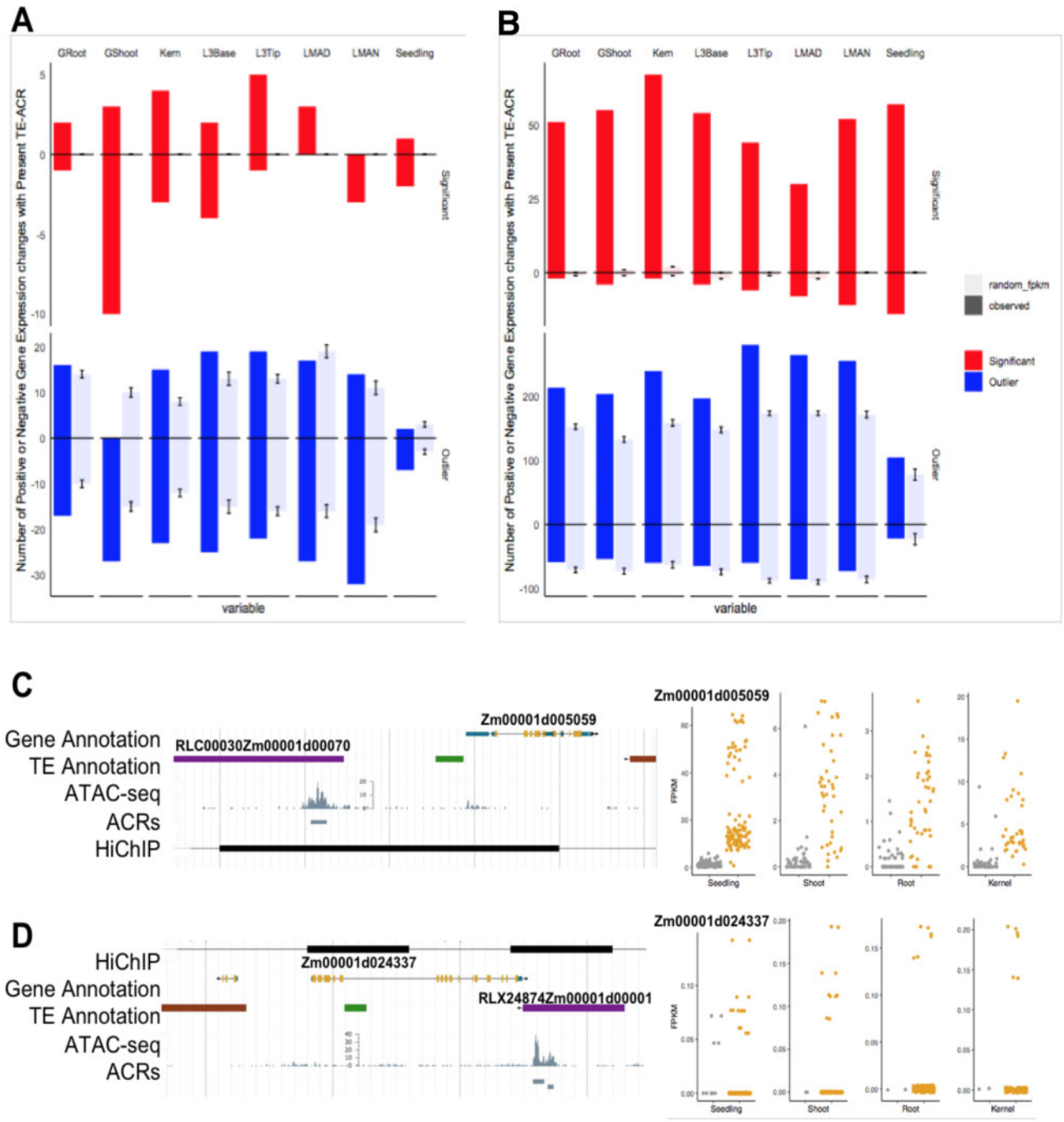

Figure 6 TE PAV association with gene expression. (A) Number of TE insertions that result in significant (red) or outlier (blue) expression changes in nearby genes by tissue for observed and randomized genotype or randomized RNA-seq controls shown by shading. (B) Number of TE-ACRs resulting in significant or outlier expression changes. (C/D) Examples of significant gene expression changes associated with TE presence. Left: genome browser view of the TE, gene and ACR. Right: dotplot of gene expression for genotypes present (yellow) or absent (gray) for seedling, shoot, root, and kernel corresponding to the TE–gene pair.

regions of accessible chromatin can potentially result in the loss of accessibility and gains of DNA methylation for the flanking sequences. We observe many examples of TE insertions into ACRs for which the regions immediately flanking the TE remain unmethylated and potentially accessible. In some cases, the insertion of a TE within a larger ACR results in two smaller ACRs on either side of the TE. Often these regions have partial overlap with the edges of the TE. However, there is a subset of examples of TE insertions into accessible regions where the previously accessible and unmethylated regions exhibit high levels of methylation on one or both sides of the TE insertion in the TE-present genotype.

TEs that introduce novel ACRs have the challenge of maintaining an unmethylated ACR within a highly targeted and condensed repetitive sequence. Even in the TEs that contain an ACR, we find that the remainder of the TE is highly methylated. When assessed across three additional genotypes, the methylation state of these ACRs was more variable than other unmethylated regions that were outside of TEs. This may suggest that the presence of a TE containing a putative regulatory element in the B73

genome may not predict the presence of an active regulatory element in other genotypes. These would result in the potential for facultative epialleles (Richards 2006; Springer and Schmitz 2017) in which some haplotypes with the TE contain an active regulatory element while others would have a silencer element. This would complicate our ability to make associations between the genetic presence/absence of the TE and the expression level of nearby genes. In our analyses, we made the assumption that when the TE is present the accessible, unmethylated region will be conserved. However, epigenetic polymorphisms would significantly reduce our power. Indeed, careful examination of some examples such as those in Figure 6, C and D, reveals that, even though the TE presence is often associated with higher expression for the nearby genes, there are some haplotypes that contain the TE but do not show high expression for the nearby gene. These may reflect epigenetic silencing of the regulatory element within these TEs. Alternatively, this could reflect potential variation in *trans*-acting factors.

TE influences on regulatory variation for genes

There are massive numbers of polymorphic TE insertions between any two maize genotypes (Wang and Dooner 2006; Springer *et al.* 2018; Sun *et al.* 2018; Anderson *et al.* 2019a). The majority of these polymorphisms likely have little or no impact on gene products or gene expression and are essentially neutral polymorphisms. However, if even a small portion influences gene expression, this could account for a major source of regulatory variation. In this study, we have used chromatin accessibility profiling to narrow the set of TE polymorphisms that might result in altered expression for nearby genes. Specifically, we focused on two classes of polymorphisms that could be assessed based on high-quality chromatin accessibility data for the B73 genome (Ricci *et al.* 2019). The presence of an ACR within a TE in B73 enables us to investigate whether the presence of this TE in other maize genotypes is associated with high, or lower, expression of the nearby gene. Alternatively, the presence of an ACR in B73 with a polymorphic TE insertion in PH207, Mo17, or W22 allows for an understanding of how the interruption of an ACR may influence gene expression.

Even in this focused set of TE polymorphisms, we find that most of the TE polymorphisms are not significantly associated with altered expression of nearby genes in the tissues we monitored. A majority of genes were found to have little to no change in the expression level relative to TE presence/absence (80% of TE-ACRs and 87% of TE insertions into ACRs). This could suggest that these TE-ACRs do not influence expression of the nearby gene. However, it is also possible that in some cases we have not examined the right tissue or growth condition, or that epigenetic instability of the ACR within TEs might complicate our ability to make a genetic association as described above. While the majority of TE polymorphisms were not significantly associated with expression for nearby genes, there are 21 examples of TE insertions into ACRs and 190 examples of TE containing ACRs that are significantly associated with the expression of nearby genes. The lack of strong effects for TE insertions into ACRs was somewhat surprising. In some cases, the TE insertions into ACRs may result in dividing a single ACR into two regions separated by the TE. This would predict that there would be instances in the B73 genome in which there are two nearby ACRs that are separated by a TE and the insertion did not necessarily disrupt the functionality of the regulatory region. Interestingly, the examples of TE containing ACRs that are significantly associated with expression are heavily biased toward examples in which the nearby gene is higher expressed. This suggests that the TE is providing an enhancer that increases gene expression. In addition to the significant associations, there are many other examples in which there is substantial variation in expression levels for haplotypes with and without the TE but which lack any statistical significance (outliers). These likely represent examples in which the haplotype with (or without) the TE is rare and only present in one or two genotypes. This might be expected in situations in which TE insertions influence expression resulting in substantial deleterious effects. These outliers are enriched for lower expression of the nearby gene for TE insertions into ACRs but higher expression for the nearby gene for TEs containing ACRs.

A key question we wrestled with in this study is whether the presence of an ACR within a TE was a property of certain TE families. Given the sequence conservation within TE families, we might predict that the presence of a regulatory element would be conserved in many members of the same TE family. Searching for this consistency is complicated by the focus on uniquely mapping reads. Indeed, we have likely greatly underestimated the number of ACRs within TEs (Supplementary Figure S5). In many cases, we would only find an ACR in one member of a multi-TE family. These might suggest that the ability to form an accessible region is attributed to both the genetic sequence of the TE and local chromatin context. We do find examples of TE families in which there are multiple members with an ACR but even in these families there are other members that lack the ACR (Supplementary Figure S7). In this analysis, we do not find strong evidence for TE families in which a common regulatory element is present and accessible for many elements of the same family. This highlights the role for both the DNA sequence of TEs and the chromatin landscape of these TEs.

Identification of ACRs across the genome has enabled us to narrow in on the ~1% of the genome with potential regulatory function (Rodgers-Melnick *et al.* 2016; Oka *et al.* 2017; Zhao *et al.* 2018; Ricci *et al.* 2019). By assessing how TE variation could contribute to polymorphisms for these accessible regions, we have characterized the potential for TEs to disrupt ACRs or contribute novel ACRs to genes. We assessed both the chromatin and regulatory consequences of these polymorphisms. We find evidence that a subset of TEs containing ACRs are likely providing enhancers to nearby genes. There was little evidence for widespread consequences of insertions of TEs into ACRs. However, many of the TE polymorphisms that strongly influence gene expression might represent rare deleterious alleles. This analysis highlights the potential for TEs to influence gene expression by creating novel expression patterns rather than simply disrupting existing information.

Acknowledgment

The Minnesota Supercomputing Institute at the University of Minnesota provided computational resources that contributed to this research.

Funding

This work was funded by NSF IOS-1934384 to N.M.S. and C.N.H, NSF IOS-1856627 and NSF IOS-1844427 to R.J.S., and NSF IOS-1546727 to C.N.H. J.M.N. is supported by a Hatch grant from the Minnesota Agricultural Experiment Station (MIN 71-068). R.J.S. is a Pew Scholar in the Biomedical Sciences, supported by The Pew Charitable Trusts. A.P.M is supported by NSF PRFB IOS-1905869.

Conflicts of interest

None declared.

Literature cited

Anderson SN, Stitzer MC, Brohammer AB, Zhou P, Noshay JM, *et al.* 2019a. Transposable elements contribute to dynamic genome content in maize. Plant J. **100**:1052–1065

Anderson SN, Stitzer MC, Zhou P, Ross-Ibarra J, Hirsch CD, *et al.* 2019b. Dynamic patterns of transcript abundance of transposable element families in maize. G3 (Bethesda). **9**:3673–3682.

Baucom RS, Estill JC, Chaparro C, Upshaw N, Jogi A, *et al.* 2009. Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. PLoS Genet. **5**:e1000732.

Bennetzen JL. 2000. Transposable element contributions to plant gene and genome evolution. Plant Mol Biol. **42**:251–269.

Bennetzen JL, Kellogg EA. 1997. Do plants have a one-way ticket to genomic obesity? Plant Cell. **9**:1509–1514.

Briggs WH, McMullen MD, Gaut BS, Doebley J. 2007. Linkage mapping of domestication loci in a large maize teosinte backcross resource. Genetics. **177**:1915–1928.

Castelletti S, Tuberosa R, Pindo M, Salvi S. 2014. A MITE transposon insertion is associated with differential methylation at the maize flowering time QTL Vgt1. G3 (Bethesda). 10.1534/g3.114.010686.

Cavrak VV, Lettner N, Jamge S, Kosarewicz A, Bayer LM, *et al.* 2014. How a retrotransposon exploits the plant's heat stress response for its activation. PLoS Genet. **10**:e1004115.

Choi JY, Purugganan MD. 2018. Evolutionary epigenomics of retrotransposon-mediated methylation spreading in rice. Mol Biol Evol. **35**:365–382. doi:10.1093/molbev/msx284

Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet. **18**:71–86.

Clark RM, Wagler TN, Quijada P, Doebley J. 2006. A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture. Nat Genet. **38**:594–597.

Dietrich CR, Cui F, Packila ML, Li J, Ashlock DA, *et al.* 2002. Maize Mu transposons are targeted to the 5' untranslated region of the gl8 gene and sequences flanking Mu target-site duplications exhibit nonrandom nucleotide composition throughout the genome. Genetics. **160**:697–716.

Doebley J, Stec A, Hubbard L. 1997. The evolution of apical dominance in maize. Nature. **386**:485–488.

Eichten SR, Ellis NA, Makarevitch I, Yeh CT, Gent JI, *et al.* 2012. Spreading of heterochromatin is limited to specific families of maize retrotransposons. PLoS Genet. **8**:e1003127.

Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. Nat Rev Genet. **9**:397–405.

Guo C, Spinelli M, Ye C, Li QQ, Liang C. 2017. Genome-wide comparative analysis of miniature inverted repeat transposable elements in 19 Arabidopsis thaliana. Sci Rep. **7**:2634.https://doi.org/10.1038/s41598-017-02855-1.

Han Y, Qin S, Wessler SR. 2013. Comparison of class 2 transposable elements at superfamily resolution reveals conserved and distinct features in cereal grass genomes. BMC Genomics. **14**:71.

Hansey CN, Johnson JM, Sekhon RS, Kaeppler SM, de Leon N. 2011. Genetic diversity of a maize association population with restricted phenology. Crop Sci. **51**:704–715.

Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, *et al.* 2014. Insights into the maize pan-genome and pan-transcriptome. Plant Cell. **26**:121–135.

Hirsch C, Hirsch CD, Brohammer AB, Bowman MJ, Soifer I, *et al.* 2016. Draft assembly of elite inbred line PH207 provides insights into genomic and transcriptome diversity in maize. Plant Cell. 2700–2714. **28**:

Hollister JD, Gaut BS. 2009. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res. **19**:1419–1428.

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR. 2004. Pack-MULE transposable elements mediate gene evolution in plants. Nature. **431**:569–573.

Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, *et al.* 2016. The complex sequence landscape of maize revealed by single molecule technologies, 1–19.

Jordan IK, Jordan IK, Rogozin IB, Glazko GV, Koonin EV. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet. **19**:68–72.

Kolkman JM, Conrad LJ, Farmer PR, Hardeman K, Ahern KR, *et al.* 2005. Distribution of activator (Ac) throughout the maize genome for use in regional mutagenesis. Genetics. **169**:981–995. doi:10.1534/genetics.104.033738.

Kremling KAG, Chen S-Y, Su M-H, Lepak NK, Romay MC, *et al.* 2018. Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. Nature. **555**:520–523.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. **25**:1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, *et al.*; 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. Bioinformatics. **25**:2078–2079.

Lisch D. 2013. How important are transposons for plant evolution? Nat Rev Genet. **14**:49–61.

Lisch D. 2015. Mutator and MULE transposons. Microbiol Spectr. **2015**; 3:MDNA3-0032. doi:10.1128/microbiolspec.MDNA3-0032-2014.

Liu S, Yeh C-T, Ji T, Ying K, Wu H, *et al.* 2009. Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. PLoS Genet. **5**:e1000733. doi:10.1371/journal.pgen.1000733.

Lowe CB, Bejerano G, Haussler D. 2007. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. Proc Natl Acad Sci USA. **104**:8005–8010.

Lu Z, Marand AP, Ricci WA, Ethridge CL, Zhang X, *et al.* 2019. The prevalence, evolution and chromatin signatures of plant regulatory elements. Nat Plants. **5**:1250–1259.

Makarevitch I, Waters AJ, West PT, Stitzer M, Hirsch CN, *et al.* 2015. Transposable elements contribute to activation of maize genes in response to abiotic stress. PLoS Genet. **11**:e1004915.

Mao H, Wang H, Liu S, Li Z, Yang X, *et al.* 2015. A transposable element in a NAC gene is associated with drought tolerance in maize seedlings. Nat. Commun. **6**:8326.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. **17**:10–12.

Maside X, Bartolomé C, Charlesworth B. 2002. S-element insertions are associated with the evolution of the Hsp70 genes in Drosophila melanogaster. Curr Biol. **12**:1686–1691.

Mazaheri M, Heckwolf M, Vaillancourt B, Gage JL, Burdo B, *et al.* 2019. Genome-wide association analysis of stalk biomass and anatomical traits in maize. BMC Plant Biol. **19**:45.

McClintock B. 1951. Chromosome organization and genic expression. Cold Spring Harb Symp Quant Biol. **16**:13–47.

Michael TP, Jackson S. 2013. The first 50 plant genomes. Plant Genome.val. 6 pg. 2

Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN, *et al.* 2009. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature. **461**:1130–1134.

Widespread natural variation of DNA methylation within angiosperms. Genome Biol. 17; 194.

Nishihara H, Smit AFA, Okada N. 2006. Functional noncoding sequences derived from SINEs in the mammalian genome. Genome Res. **16**:864–874.

Noshay JM, Anderson SN, Zhou P, Ji L, Ricci W, *et al.* 2019. Monitoring the interplay between transposable element families and DNA methylation in maize. PLoS Genet. **15**:e1008291.

O'Connor CH, Qiu Y, Coletta RD, Monnahan PJ, *et al.* 2020. Population level variation of transposable elements in a maize diversity panel. BioRxiv. 10.1101/2020.09.25.314401.

Oka R, Zicola J, Weber B, Anderson SN, Hodgman C, *et al.* 2017. Genome-wide mapping of transcriptional enhancer candidates using DNA and chromatin features in maize. Genome Biol. **18**: 137.

"Picard Toolkit." 2018. Broad Institute, GitHub Repository. http://broadinstitute.github.io/picard/; Broad Institute.

Ricci WA, Lu Z, Ji L, Marand AP, Ethridge CL, *et al.* 2019. Widespread long-range cis-regulatory elements in the maize genome. Nat Plants. **5**:1237–1249.

Richards EJ. 2006. Inherited epigenetic variation—revisiting soft inheritance. Nat Rev Genet. **7**:395–401.

Rodgers-Melnick E, Vera DL, Bass HW, Buckler ES. 2016. Open chromatin reveals the functional maize genome. Proc Natl Acad Sci USA. **113**:E3177–84.

Schnable PS, Ware D, Fulton RS, Stein JC, Pasternak S, *et al.* 2009. The B73 maize genome: complexity, diversity, and dynamics. Science. **326**:1112–1115.

Sheffield NC, Thurman RE, Song L, Safi A, Stamatoyannopoulos JA, *et al.* 2013. Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. Genome Res. **23**:777–788.

Springer NM, Anderson SN, Andorf CM, Ahern KR, Bai F, *et al.* 2018. The maize W22 genome provides a foundation for functional genomics and transposon biology. Nat Genet. https://doi.org/10.1038/s41588-018-0158-0.

Springer NM, Lisch D, Li Q. 2016. Creating order from chaos: epigenome dynamics in plants with complex genomes. Plant Cell. **28**: 314–325.

Springer NM, Schmitz RJ. 2017. Exploiting induced and natural epigenetic variation for crop improvement. Nat Rev Genet. https://doi.org/10.1038/nrg.2017.45.

Stitzer MC, Anderson, SN Springer, NM Ross-Ibarra J. 2019. The genomic ecosystem of transposable elements in maize. BioRxiv. 10.1101/559922.

Studer A, Zhao Q, Ross-Ibarra J, Doebley J. 2011. Identification of a functional transposon insertion in the maize domestication gene tb1. Nat Genet. **43**:1160–1163.

Sun S, Zhou Y, Chen J, Shi J, Zhao H, *et al.* 2018. Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. Nat Genet. **50**:1289–1295.

Vollbrecht E, Duvick J, Schares JP, Ahern KR, Deewatthanawong P, *et al.* 2010. Genome-wide distribution of transposed dissociation elements in maize. Plant Cell. **22**:1667–1685. doi: 10.1105/tpc.109.073452.

Wang Q, Dooner HK. 2006. Remarkable variation in maize genome structure inferred from haplotype diversity at the bz locus. Proc Natl Acad Sci USA. **103**:17644–17649.

Weil C, Martienssen R. 2008. Epigenetic interactions between transposons and genes: lessons from plants. Curr Opin Genet Dev. **18**: 188–192.

Wessler S., Bureau TE, White, SE. 1995. LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. Curr Opin Genet Dev. **5**:814–821.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, *et al.* 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. **8**:973–982.,

Wyler M, Stritt C, Walser JC, Baroux C, Roulin AC. 2020. Impact of transposable elements on methylation and gene expression across natural accessions of Brachypodium distachyon. Genome Biol Evol. **12**: 1994–2001.

Xi Y, Li W. 2009. BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics. **10**:232.

Yang Q, Li Z, Li W, Ku L, Wang C, *et al.* 2013. CACTA-like transposable element in ZmCCT attenuated photoperiod sensitivity and accelerated the postdomestication spread of maize. Proc Natl Acad Sci USA. **110**:16969–16974.

Zerjal T, Rousselet A, Mhiri C, Combes V, Madur D, *et al.* 2012. Maize genetic diversity and association mapping using transposable element insertion polymorphisms. Theor Appl Genet. **124**:1521–1537.

Zhang P, Allen WB, Nagasawa N, Ching AS, Heppard EP, *et al.* 2012. A transposable element insertion within ZmGE2 gene is associated with increase in embryo to endosperm ratio in maize. Theor Appl Genet. **125**:1463–1471.

Zhao H, Zhang W, Chen L, Wang L, Marand AP, *et al.* 2018. Proliferation of regulatory DNA elements derived from transposable elements in the maize genome. Plant Physiol. **176**: 2789–2803. doi:10.1104/pp.17.01467.

Zhou P, Hirsch CN, Briggs SP, Springer NM. 2019. Dynamic patterns of gene expression additivity and regulatory variation throughout maize development. Mol Plant. **12**:410–425.

Communicating editor: K. Bomblies