





# Stability of DNA methylation and chromatin accessibility in structurally diverse maize genomes

Jaclyn M. Noshay <sup>1</sup>, Zhikai Liang,<sup>1</sup> Peng Zhou,<sup>1</sup> Peter A. Crisp,<sup>2</sup> Alexandre P. Marand,<sup>3</sup> Candice N. Hirsch <sup>4</sup>, Robert J. Schmitz <sup>3</sup> and Nathan M. Springer <sup>1,\*</sup>

<sup>1</sup>Department of Plant and Microbial Biology, University of Minnesota, Saint Paul, MN 55108, USA

<sup>2</sup>School of Agriculture and Food Sciences, University of Queensland, St Lucia, QLD 4072, Australia

<sup>3</sup>Department of Genetics, University of Georgia, Athens, GA 30602, USA

<sup>4</sup>Department of Agronomy and Plant Genetics, University of Minnesota, Saint Paul, MN 55108, USA

\*Corresponding author: Email: [springer@umn.edu](mailto:springer@umn.edu)

## Abstract

Accessible chromatin and unmethylated DNA are associated with many genes and cis-regulatory elements. Attempts to understand natural variation for accessible chromatin regions (ACRs) and unmethylated regions (UMRs) often rely upon alignments to a single reference genome. This limits the ability to assess regions that are absent in the reference genome assembly and monitor how nearby structural variants influence variation in chromatin state. In this study, *de novo* genome assemblies for four maize inbreds (B73, Mo17, Oh43, and W22) are utilized to assess chromatin accessibility and DNA methylation patterns in a pan-genome context. A more complete set of UMRs and ACRs can be identified when chromatin data are aligned to the matched genome rather than a single reference genome. While there are UMRs and ACRs present within genomic regions that are not shared between genotypes, these features are 6- to 12-fold enriched within regions between genomes. Characterization of UMRs present within shared genomic regions reveals that most UMRs maintain the unmethylated state in other genotypes with only ~5% being polymorphic between genotypes. However, the majority (71%) of UMRs that are shared between genotypes only exhibit partial overlaps suggesting that the boundaries between methylated and unmethylated DNA are dynamic. This instability is not solely due to sequence variation as these partially overlapping UMRs are frequently found within genomic regions that lack sequence variation. The ability to compare chromatin properties among individuals with structural variation enables pan-epigenome analyses to study the sources of variation for accessible chromatin and unmethylated DNA.

**Keywords:** DNA methylation; chromatin accessibility; comparative epigenomics

## Introduction

The 2.1Gb maize B73 genome was first assembled in 2009 and contains ~80% repetitive sequence (Schnable et al. 2009). Unlike model species such as *Arabidopsis thaliana*, maize has transposable elements and highly methylated regions that are interspersed with genic regions of the genome (The Arabidopsis Genome Initiative 2000; Baucom et al. 2009; Springer and Schmitz 2017). One challenge in complex crop genomes such as maize is the identification of regulatory elements within genomes. There are opportunities to utilize both chromatin properties such as DNA methylation or chromatin accessibility to identify functional elements (Crisp et al. 2020).

The maize genome is highly methylated and regions containing DNA methylation can be sub-classified based on the specific sequence context of the methylation. High levels of CG and CHG (H = A, C, or T) methylation without CHH methylation are often found over transposable elements and other repetitive regions of the maize genome, while CG-only methylation is observed frequently within gene bodies (West et al. 2014; Niederhuth et al. 2016; Crisp et al. 2020). CHH methylation, which is largely the

result of RNA-directed DNA methylation (RdDM), is found near highly expressed genes (Gent et al. 2013; Li et al. 2015a; Niederhuth et al. 2016). A small proportion of the maize genome lacks DNA methylation in any sequence context and these unmethylated regions (UMRs) likely reflect regions with potential roles in regulation of gene expression (Oka et al. 2017; Ricci et al. 2019; Crisp et al. 2020; Hoefsloot and Stam 2020).

Chromatin accessibility is another feature of chromatin that can be used to identify genomic regions with roles in regulation of transcription. In maize, ~1% of the genome contains accessible chromatin when profiled with a single tissue type (Rodgers-Melnick et al. 2016). Profiles of chromatin accessibility combined with other chromatin modifications have identified potential regulatory elements in the maize genome (Oka et al. 2017; Ricci et al. 2019). While chromatin accessibility is quite useful for identifying regulatory elements in a particular tissue, this property is highly dynamic with changes between tissue types or cells (Ricci et al. 2019; Crisp et al. 2020; Marand et al. 2020). The vast majority of accessible chromatin occurs in UMRs of the genome. However, there are additional UMRs that do not exhibit chromatin accessibility. These likely reflect the fact that the UMRs of the

Received: March 10, 2021. Accepted: May 27, 2021

© The Author(s) 2021. Published by Oxford University Press on behalf of Genetics Society of America.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

genome are quite stable in vegetative tissues while chromatin accessibility is highly tissue-specific (Schmitz *et al.* 2013; Kawakatsu *et al.* 2016; Crisp *et al.* 2020; Marand *et al.* 2020). To date, the analysis of chromatin accessibility in maize has largely focused on the accessible regions within the B73 genome.

The analysis of chromatin properties within the B73 reference genome has been useful for functional annotation of the genome. However, there is also value in assessing natural variation for the chromatin properties in different inbred lines of maize. While chromatin accessibility studies have largely focused on B73, many studies have compared DNA methylation between maize genotypes (Eichten *et al.* 2013; Regulski *et al.* 2013; Li *et al.* 2015b; Anderson *et al.* 2018; Xu *et al.* 2019, 2020). These studies have found many examples of DNA methylation variation. Changes in DNA methylation can occur due to alterations in DNA sequence such as transposon insertions (Noshay *et al.* 2019) or can occur in regions with no genetic changes (Eichten *et al.* 2011). The ability to fully compare DNA methylation patterns among genotypes and to investigate the role of structural variation has been limited due to reliance upon a single reference genome for comparisons.

The genome content varies substantially among maize genotypes (Fu and Dooner 2002; Springer *et al.* 2009; Swanson-Wagner *et al.* 2010; Anderson *et al.* 2019; Hufford *et al.* 2021). The availability of multiple *de novo* assembled reference genomes has enabled whole-genome comparisons of genome content (Hirsch *et al.* 2016; Springer *et al.* 2018; Sun *et al.* 2018; Haberer *et al.* 2020; Hufford *et al.* 2021). Many of the sequences present in any one inbred are not present at collinear regions in other genomes (Fu and Dooner 2002; Sun *et al.* 2018; Haberer *et al.* 2020). This results in a pan-genome that contains more genes and transposons than any individual maize inbred (Hirsch *et al.* 2014; Anderson *et al.* 2019; Hufford *et al.* 2021). While there is evidence for genome content variation among maize inbreds it has been difficult to assess the chromatin of the pan-genome due to technical difficulties in connecting the same sequence regions between genotypes.

In this study, we generated DNA methylation and chromatin accessibility profiles from four maize inbreds that each have *de novo* genome assemblies. UMRs and ACRs are identified for each genotype based on alignment of the chromatin data to the B73v4 genome and the genome from which it was generated. Chromosomal alignments were used to classify shared and non-shared sequences between genomes. UMRs and ACRs are substantially depleted within the nonshared portions of the genome. Within the shared regions of the genome we assessed the stability of UMRs between genotypes. While the majority of UMRs in these regions have an overlapping UMR in another genotype, over 50% do not have identical coordinates due to shifts in the precise boundaries between methylated and unmethylated DNA. These UMRs with shifted boundaries account for a large portion of the differentially methylated regions between two genotypes. The partially overlapping UMRs are not enriched for variable chromatin accessibility or changes in expression of nearby genes, suggesting that differences in the specific boundaries between methylated and unmethylated DNA are tolerated with little functional impact.

## Materials and methods

### Reference genomes

Whole-genome assemblies for four maize inbred lines, B73 (Jiao *et al.* 2017), W22 (Springer *et al.* 2018), Mo17 (Sun *et al.* 2018), and Oh43 (Hufford *et al.* 2021) were used for genome-wide analyses.

All analyses were performed on assemblies of chromosomes 1–10 while all unplaced scaffolds were disregarded due to the inability to compare these regions across genotypes. Filtered gene and structural TE annotations (Stitzer *et al.*; Anderson *et al.* 2019) were used.

### Sample collection

Maize B73, W22, Mo17, and Oh43 plants were grown under 16 hours/8 hours 30°C/20°C day/night for 13 days in the growth chamber of the University of Minnesota. DNA was extracted from leaves of 2-week old V2 plants using the DNeasy Plant Mini kit (Qiagen). Four or five biological replicates consisting of a pool of tissue from 4 plants were collected for each genotype. Two of these biological replicates were sampled for profiling of DNA methylation and chromatin accessibility while all biological replicates were used for RNAseq.

### WGBS protocol

Two biological replicates of each genotype (B73, Mo17, W22, and Oh43) were generated. 1 µg of DNA in 50 µg of water was sheared using an Ultrasonicator to approximately 200–350 bp fragments. Twenty microliter of sheared DNA was then bisulfite converted using the EX DNA Methylation-Lightning Kit (Zymo Research) as per the manufacturer's instructions and eluted in a final volume of 15 µl. Then 7.5 µl of the fragmented bisulfite-converted sample was used as input for library preparation using the ACCEL-NGS Methyl-Seq DNA Library Kit (SWIFT Biosciences). Library preparation was performed as per the manufacturer's instructions. The indexing PCR was performed for 5 cycles. Libraries were then pooled and sequenced on a NovaSeq 6000 in high output mode 125 bp paired end reads over a single lane at the University of Minnesota Genomics Center. WGBS data generated in this study are deposited at NCBI SRA and available under accession.

Trim\_galore (Martin 2011) was used to trim adapter sequences and read quality was assessed with the default parameters in paired-end read mode plus a hard clip of 20 bp on each read due to SWIFT protocol specifications. Reads that passed quality control were aligned to their corresponding genome assemblies. Alignments were conducted using BSMAP-2.90 (Xi and Li 2009), allowing only unique hits with 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") 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 methylation level for each cytosine using BSMAP tools.

### Methylation data summary

Methylation levels were summarized using the bsmmap methratio.py script to group by context (CG, CHG, and CHH). The number of cytosines in every 100 bp bin of the genome was determined and the proportion of cytosines defined as methylated was calculated. Coverage was calculated as CT/# of sites for each context. Methylation domain was classified for each 100 bp bin based on the protocol described in Crisp *et al.* (2020).

Briefly, each 100 bp bin of the genome was classified into one of six methylation domains ("missing data," "RdDM," "heterochromatin," "CG-only," "unmethylated," or "intermediate,"). Tiles were classified in a hierarchical order first by defining any tiles with less than two cytosines or less than 5x coverage as missing data. Remaining tiles were defined by the level of methylation; RdDM if CHH methylation was greater than

15%; heterochromatin if CG and CHG methylation was 40% or greater; CG-only if CG methylation was greater than 40%; unmethylated if CG, CHG, and CHH were less than 10%; and intermediate if methylation was 10% or greater but less than 40%. UMRs were defined by grouping adjacent unmethylated bins or missing data (as long as the resulting UMR contained <33% missing data) and all UMRs less than 300 bp were removed.

UMRs were classified relative to annotated genes as described in Ricci et al. (2019). All UMRs that overlap a gene were first defined as genic. Nongenic UMRs were further classified as gene-proximal if they were within 2000 bp. All remaining UMRs that do not overlap any sequence within 2000 bp of the annotated gene are classified as intergenic.

## ATAC-seq protocol and ACR classification

ATAC-seq libraries were generated as described in Lu et al. (2017). Two biological replicates of each genotype (B73, Mo17, W22, and Oh43) were generated from the same samples as those used for WGBS data generation. Raw reads per sample were preprocessed with Trim\_galore. Trimmed reads were aligned to the *Zea mays* B73v4 genome and the genome assembly specific to each sample using Bowtie v1.2.3 with the following parameters: “bowtie -X 1000 -m 1 -v 2 -best -strata.” Aligned reads were converted to bam files and sorted using SAMtools v1.9. Clonal duplicates were removed using Picard MarkDuplicates v2.23.3 (<http://broadinstitute.github.io/picard/>). Input data of maize B73 was retrieved from a previous publication and processed to obtain bam files with clonal duplicates removed. MACS2 was employed to call initial ACRs with Input data as control (-c) and sample data as treatment (-t) using the following parameter “-g 2.1e9 -keep-dup all -nomodel -extsize 147.” The post-processing followed the same procedure as a prior publication (Ricci et al. 2019) to produce high-confidence ACRs. Specifically, (1) Initial ACRs were split into 50 bp windows with 25 bp steps; (2) the Tn5 integration frequency in each window was calculated and normalized to the average frequency in the total genome; (3) windows with the normalized frequency greater than 25 were merged together allowing 150 bp gaps; (4) only merged regions greater than 50 bp were retained; (5) the mitochondrial or chloroplast genome from NCBI Organelle Genome Resources were removed using blast against sequences within merged ACR regions. The sites within ACRs that had the highest Tn5 integration frequency were defined as summits.

## RNA-seq protocol

RNA-seq data were generated in 150 bp paired-end mode using NovaSeq 6000. B73, W22, and Mo17 reads were retrieved from the NCBI SRA accession PRJNA657262 (Liang et al. 2021) and Oh43 reads were deposited into NCBI SRA accession PRJNA692023. All of the raw reads were preprocessed using Trim\_galore and aligned against the B73 AGPv4 reference genome using HISAT2 v2.1.0 (Kim et al. 2015). Gene annotations and disjointed TE annotations were used as described above. Gene exon regions were subtracted from TE regions and then appended to the original TE annotation to remove ambiguous mapping between genes and TEs. Reads per gene or TE was determined using HTSeq-count v0.11.2 (Anders et al. 2015) and raw count data was input into DESeq2 (Love et al. 2014) to identify differentially expressed genes or TE elements.

The mean value for each feature (gene or TE) was calculated from 4 or 5 replicates. Any feature with a mean value greater than 1 was considered “expressed.” UMRs were associated with genes and TEs based on location relative to the feature. B73 UMRs which overlapped the annotated sequence coordinates

within the genome being assessed were classified as “genic” or “TE.” Those not overlapping a gene but within 2 kb of the gene start or end were classified as “proximal.”

## Cross-genotype mapping

Genome sequence from Mo17, W22, and Oh43 was first aligned to the B73 reference (Jiao et al. 2017) using minimap2 (Li 2018). The resulting alignments were merged and cleaned (removing overlapping alignment blocks and alignment blocks containing assembly gaps) using in-house perl scripts. BLAT Chain/Net tools were then used to create a single coverage best alignment net between the query genome (one of Mo17, W22, and Oh43) and the target genome (B73). Finally, a genome-wide synteny chain file was built for each genotype (against B73), enabling downstream analyses such as variant detection and 100-bp tile liftover. Alignment pipeline and scripts are available on Github (<https://github.com/baudisgroup/segment-liftover>). The sequence was extracted for all 100 bp bins in the B73 genome and aligned to Mo17, W22, and Oh43. Each bin was determined to be unmappable or mappable. Mappable bins were assigned coordinates in the nonB73 genome. The number of single nucleotide polymorphisms and insertion/deletions for each bin was calculated. Across all genotypes, only 4% of bins were found to have  $\geq 1$  insertion/deletion and 13% contained  $\geq 1$  single nucleotide polymorphism. Bins with no more than 4 insertion/deletions of 20 bp in size were kept for analyses of shared space. Each 100 bp bin in B73 was designated as unmapped or provided matching sequence coordinates in each of the 3 other genotypes (Mo17, W22, and Oh43).

## Characterization of IBS regions

Identical by sequence regions were characterized as in Anderson et al. (2019). Briefly, SNPs between B73 and the other three genomes were identified by first aligning these genomes using minimap2 (Li 2018). BLAT (Kent 2002) chain/net tools were then used to process alignment results and build synteny chains and nets. Final SNP and InDel calling was done using Bcftools (Li 2011). SNP density for each 1 Mb bin was determined by dividing the total number of SNPs in the window by the number of base pairs in syntenic alignments in the window. Regions with SNP density lower than 0.0005 over at least a 5 Mb window were defined as IBS regions. For each comparison between B73 and a contrasting genome (W22, Mo17, or PH207), the inferred coordinates for the outermost shared site-defined B73 TEs completely within each IBS block were used to mark the boundary of the IBS region in the contrasting genome.

## Differentially methylated tiles

WGBS data aligned to the respective genome and summarized in the B73-based 100 bp coordinate system was used. Tiles were subset to those with sequence mappability and coverage in both genotypes for each pairwise comparison. Differentially methylated tiles (DMTs) were defined by a difference of 40% with at least one genotype having <10% and >40% methylation for CG and CHG contexts. CHH DMTs were defined by one genotype with <5% and >25% methylation in the 100 bp tile. DMTs in each context were determined for Mo17, W22, and Oh43 compared to B73.

## Classification of UMR variability

B73 UMRs that were mappable to sequence in another genotype were further defined by methylation state in the corresponding genome. All 100 bp bins within a defined UMR were assessed for the matching sequence coordinates in Mo17, W22, and Oh43. For



each UMR, the proportion of bins classified as methylated (including CG, CG/CHG, and CHH methylation domains) was calculated. UMRs with >50% of the bins being methylated were defined as “polymorphic UMRs” for the difference in methylation state from unmethylated in B73 to methylated in the nonB73 genotype. All other UMRs, showing an unmethylated state in both B73 and the nonB73 genotype assessed, were defined as “overlapping UMRs.”

B73 UMRs that are methylated in another genotype (polymorphic UMRs) were further classified by the type of methylation observed in the nonB73 genotype. The polymorphic UMRs were summarized by domain. The proportion of 100bp bins with a methylated domain, within the defined B73 UMR, for each methylation context was determined. Any UMR that had >50% of its methylated bins classified as a specific methylation context was declared to be variable in that context. Classification was determined first by CHH methylation, followed by CG/CHG methylation and lastly CG only methylation. Variable methylation type was defined individually for each genome based on the sequence coordinates of the B73 UMR.

B73 UMRs that are unmethylated in another genotype (overlapping UMRs) were further classified by the coordinates of the defined UMR between B73 and the nonB73 genotype. The UMRs, defined by alignment of WGBS data to the B73 reference genome, were determined and their coordinates were assessed. Pairwise comparisons were done between B73 and nonB73 genotypes. B73 UMRs that had identical 100bp bin boundaries for the defined UMR were classified as identical UMRs. B73 UMRs that had variable boundaries were classified as partial UMRs (the coordinates of the smaller UMR were maintained within the larger UMR coordinates or the coordinates are shifted and have uniquely defined unmethylated bins in each genotype).

### Classification of ACR variability

Every B73 UMR was classified based on the accessibility of that shared sequence region within B73, Mo17, W22, and Oh43. All UMRs in B73 were defined as accessible (aUMR) or inaccessible (iUMR) based on its overlap with an accessible chromatin region (ACR) in the B73 sample. For B73 aUMRs, the presence of an accessible region in the nonB73 genotypes was determined. The B73-based coordinates of the UMR in the corresponding genome were used to identify overlap with the ACRs defined in that genome. UMRs that overlap both an ACR in B73 and nonB73 genome were defined as stable ACRs. If the aUMR in B73 lacked accessibility in the nonB73 genome it was defined as B73-only ACR. Alternatively, if a UMR was inaccessible in B73 it could never be found accessible or show accessibility in the other genotype. If the iUMR lacked accessibility in the nonB73 genome, it was determined to have no ACR. If the sequence of the iUMR overlapped a defined ACR in the other genome, it was defined as a nonB73 ACR such that it was inaccessible in the B73 UMR but accessible in the shared sequence of Mo17, W22, or Oh43. The ACRs which were defined as either B73-only or nonB73-only were verified by assessing the 100bp cpm values within that region across the two genotypes.

### Data availability

Accessible chromatin data (ATAC-seq) generated for this study is available at NCBI short read archive (SRA) under accession number PRJNA709664. In this study, we also utilize previously published RNA-seq datasets that are available under accession numbers PRJNA657262 and PRJNA692023 and whole-genome bisulfite datasets that are available under accession number

PRJNA657677. Supplementary Material is available at figshare: <https://doi.org/10.25387/g3.14637411>.

## Results

### Characterization of unmethylated DNA and accessible chromatin in four maize genomes

DNA methylation (profiled using whole-genome bisulfite sequencing—WGBS, Cokus *et al.* 2008; Lister and Ecker 2009), chromatin accessibility (profiled using Assay for Transposase Accessible Chromatin-sequencing—ATAC-seq, Buenrostro *et al.* 2013), and gene expression (RNA-seq) data were generated for the same tissue sample from seedling leaf of four maize inbreds (B73, Mo17, W22, and Oh43) (Supplementary Tables S1 and S2). For all genotypes the resulting datasets were aligned to their own genome assembly and nonB73 genotypes were additionally aligned to the B73v4 reference genome assembly.

The alignment rates for the WGBS datasets were substantially higher when mapped to their respective genome assembly (~60%) compared to nonB73 samples mapped to the B73 reference genome assembly (~43%) (Supplementary Table S1). The reduced mapping rate when aligning data from nonB73 genotypes to the B73 genome assembly is likely due to polymorphisms and structural variants present between inbreds. We focused on analysis of methylation classifications based on merged replicates, since the data from the two biological replicates was highly correlated and the UMRs identified within individual samples were frequently (>97%) found in the merged sample (Supplementary Table S3). The WGBS data was used to classify the methylation state for each 100bp bin based on context-specific DNA methylation (Supplementary Figure S1A) as described previously (Crisp *et al.* 2020). Bins were classified as CHH (CHH > 15%), CG/CHG (>40% both CG and CHG), CG only (>40% CG), unmethylated (<15% CHH and <20% CG and CHG), missing data, missing sites or intermediate methylation (Supplementary Figure S1A). The majority (71–74%) of the maize genome is classified as methylated with most of this exhibiting CG/CHG methylation and in rare cases CHH methylation (Supplementary Figure S1A). A much smaller proportion (6–7%) of the genome is classified as unmethylated (Supplementary Figure S1A). In each genome, roughly 15% of the bins are classified as missing data, likely due to an inability to align WGBS reads uniquely to repetitive regions. However, the proportion of bins with missing data was substantially larger when nonB73 WGBS data were aligned to the B73 genome (Supplementary Figure S1B).

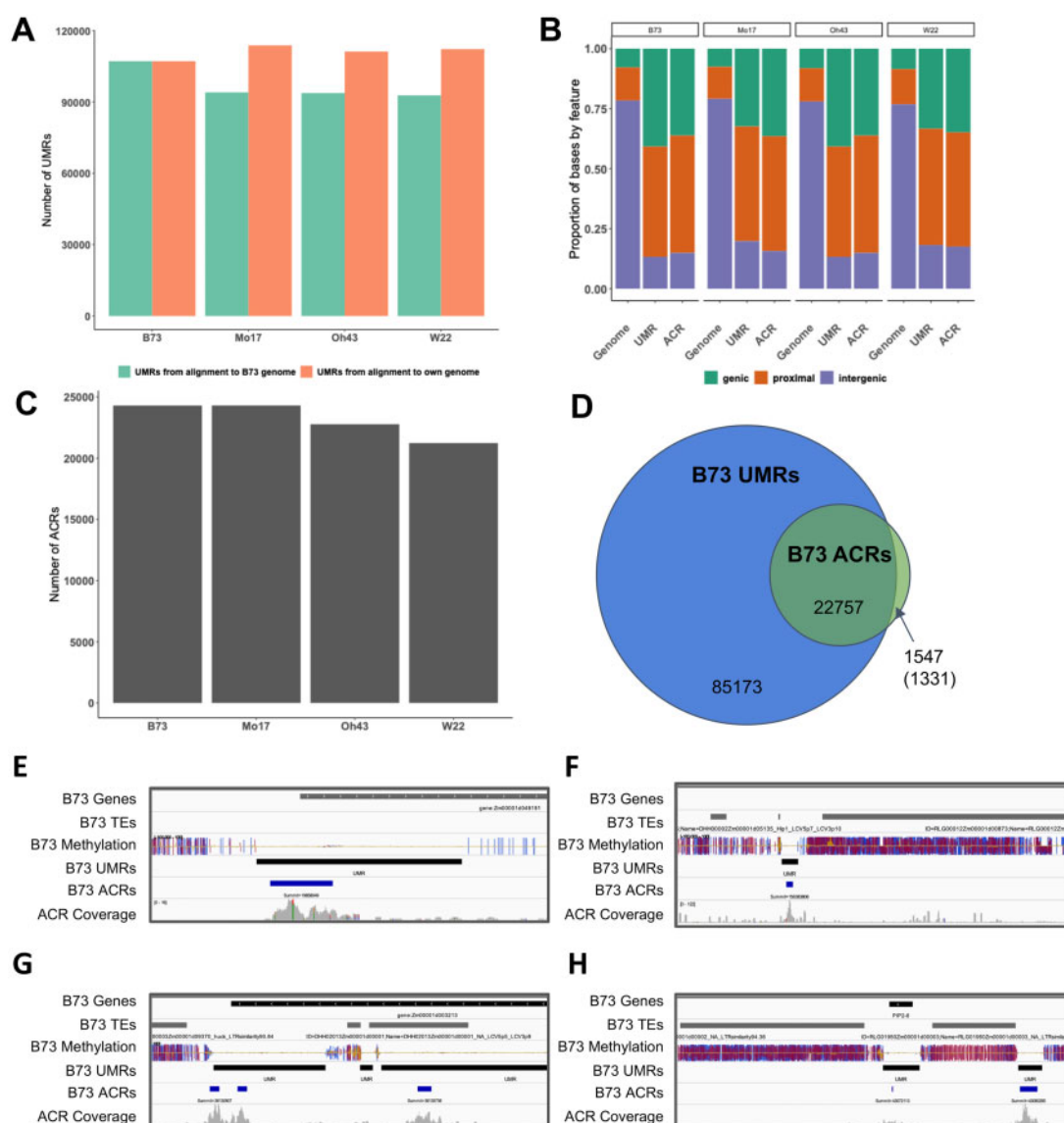
The unmethylated 100bp bins were merged and filtered (Crisp *et al.* 2020) to identify UMRs (Table 1). UMRs were defined for each inbred based on alignment to their respective genome assembly and alignment to B73 (Figure 1A). The total number of UMRs was similar across all four genotypes, although a greater number of UMRs were defined when mapping WGBS reads to the cognate genome assembly (Figure 1A). UMRs were classified as genic, proximal (<2kb from nearest gene) and intergenic (>2kb from the nearest gene) in all four genotypes based on alignment of samples to their cognate reference genome assembly. The distribution of UMRs and ACRs relative to gene annotations is fairly consistent across genotypes with >80% of UMRs being observed in genic or gene proximal regions and <20% in intergenic regions (Figure 1B).

Prior studies have found that unmethylated portions of the maize genome often contain cis-regulatory regions (Oka *et al.* 2017; Ricci *et al.* 2019; Crisp *et al.* 2020). To determine the concordance between UMRs and ACRs, we implemented ATAC-seq in

**Table 1** UMR and ACR summary statistics

Sample genotype	Reference genotype	No of bins defined as missing data	No of bins defined as Methylated <sup>a</sup>	No of bins defined as Unmethylated	No of UMRs	No of ACRs
B73	B73v4	3,511,785 (16.7%)	15,064,391 (71.5%)	1,325,187 (6.3%)	107,178	24,304
Mo17	Mo17	3,649,729 (16.6%)	15,566,698 (70.6%)	1,385,916 (6.3%)	113,838	24,309
Oh43	Oh43	3,096,596 (14.6%)	15,719,767 (74.3%)	1,445,686 (6.8%)	111,261	22,774
W22	W22	3,322,802 (15.6%)	15,315,985 (71.8%)	1,369,207 (6.4%)	112,253	21,232

<sup>a</sup> Methylated is the combined value of bins defined as CG only, CG/CHG, and CHH. Percentage is shown in ().



**Figure 1** Identification of UMRs and ACRs in maize genotypes. (A) The number of UMRs defined based on samples aligned to B73v4 (green) and their own genome assembly (orange). (B) The location of UMRs and ACRs in the genome based on gene annotations was classified as overlapping genes (green), within 2 kb of a gene (orange) and >2 kb from a gene (purple). (C) The number of ACRs defined based on the merged replicates for each genotype aligned to their respective genome assemblies. (D) Overlap between the B73 UMRs and ACRs defined based on alignments to the B73v4 genome. The number in parentheses indicates ACRs that are defined as methylated as opposed to missing data. (E–H) Accessibility is often present only for a portion of the UMR. Several B73 UMRs are shown along with ATAC-seq data. IGV (Robinson et al. 2011) snapshots of the B73 genome showing ACRs within UMR space. Tracks include B73 gene and TE annotations, B73 methylation per cytosine in all contexts (CG: blue, CHG: red, CHH: yellow), B73 UMRs (black), B73 ACRs (blue), and B73 ATAC-seq coverage (grey).

the same four genotypes. ACRs were identified in each individual sample as well as from merged biological replicates (Supplementary Table S2). We focused on analysis of the ACRs identified from the merged replicates, since the data from the two biological replicates was highly correlated ( $R^2 > 0.95$  for all genotypes—Supplementary Table S3) and the ACRs identified within individual samples were frequently found in the merged sample (Supplementary Table S3 and Figure S2). There are 21,232–24,309 ACRs present in each of the four genotypes (Table 1 and Figure 1C). Similar to the UMRs, ACRs are frequently found in genic or gene proximal regions, but 14–18% of the ACRs are found in intergenic regions >2kb from the nearest gene (Figure 1B). The vast majority of ACRs are found within UMRs in each of the four genotypes (Figure 1D and Supplementary Figure S3). While the vast majority of ACRs occur within UMRs, there are many UMRs without accessible chromatin (Figure 1E). This allows the classification of UMRs as accessible UMRs (aUMRs) or inaccessible UMRs (iUMRs) based on whether they overlap an ACR. The presence of an aUMR, which includes the presence of an ACR, is much more common within or near genes that are highly expressed, but is quite rare for lowly expressed genes (Supplementary Figure S3D). In contrast, iUMRs are present near genes with low and high expression levels, but are depleted near silent genes (Supplementary Figure S3E). While the aUMRs represent an overlap between an UMRs and chromatin accessibility, the boundaries of these regions are often not the same. The majority (97.3%) of cases represent a larger UMRs in which the ACR only covers a portion of the UMR and the ACR is often found in the center of the UMR (examples in Figure 1, E–H). This suggests that the transition from accessible to inaccessible chromatin and from unmethylated DNA to methylated DNA does not occur at the same region.

### Classification of shared and nonshared genomic regions

Previous studies have assessed natural variation in DNA methylation based on alignment to a single reference genome (Regulski et al. 2013; Li et al. 2015b). However, when WGBS data from nonB73 genotypes are aligned to the B73 genome, the proportion of regions with missing data increases substantially (Supplementary Figure S1B), and the methylation levels for genomic regions missing in B73 are not assessed. The availability of multiple reference genomes provides the opportunity to assess DNA methylation levels in the pan-genome that includes both shared (syntenic) regions of the genome with or without allelic variation, as well as nonshared regions that are present in one line and missing in another. The alignment of WGBS or ATAC-seq data to their respective genome provides the advantage of more complete characterization of DNA methylation and/or chromatin accessibility, but introduces complications for the direct comparison of specific regions among genomes.

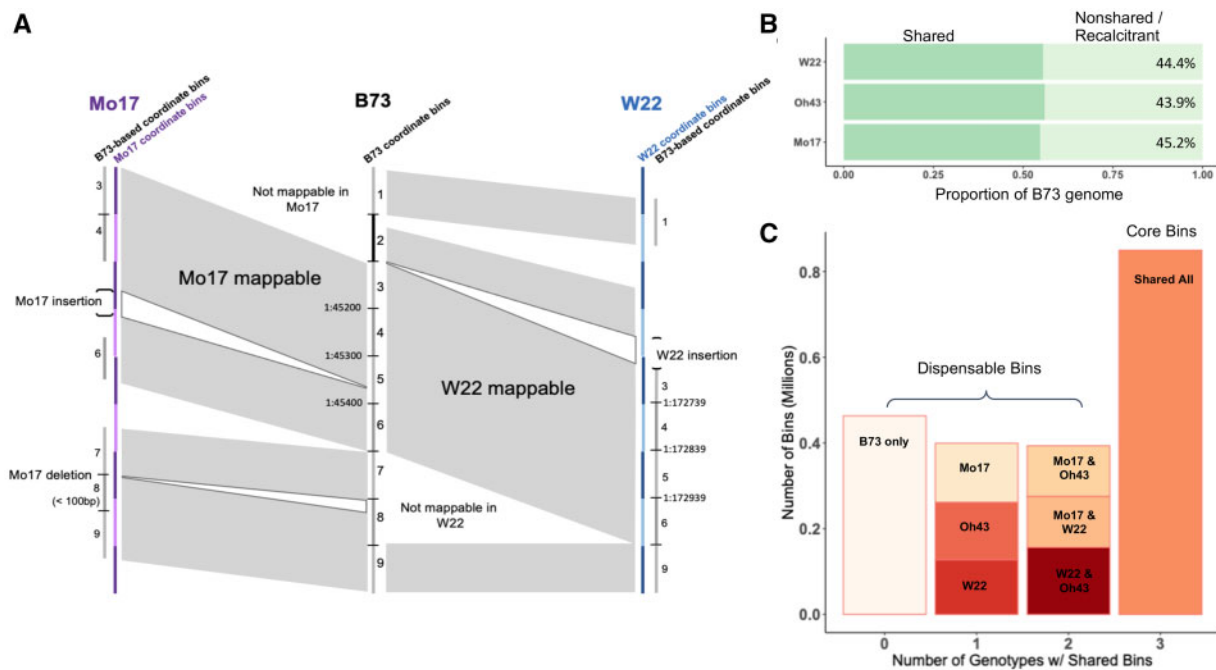
To address this complication in comparing regions across genomes, chromosomal alignments were performed between the B73 genome and the other reference genomes to identify the shared and nonshared genomic segments between any two genotypes (see Methods) (Figure 2A). The approach that was implemented employed relatively stringent criteria for identification of shared regions. The regions classified as nonshared include both structural variants and highly polymorphic regions as well as highly repetitive regions that could not be uniquely mapped. Approximately 55% of the nonB73 genome sequences could be classified as syntenic and mappable relative to B73, with the remaining 45% not aligning to the B73 genome due to

nonsyntenic sequence or unmappable regions (Figure 2B). As a quality control measure, we assessed the proportion of space classified as shared or nonshared within identity-by-state (IBS) regions between genomes. The majority (94%) of IBS regions are classified as shared between any two genomes (Supplementary Table S4) and the regions that are not classified as shared within IBS regions are highly enriched for repetitive sequences.

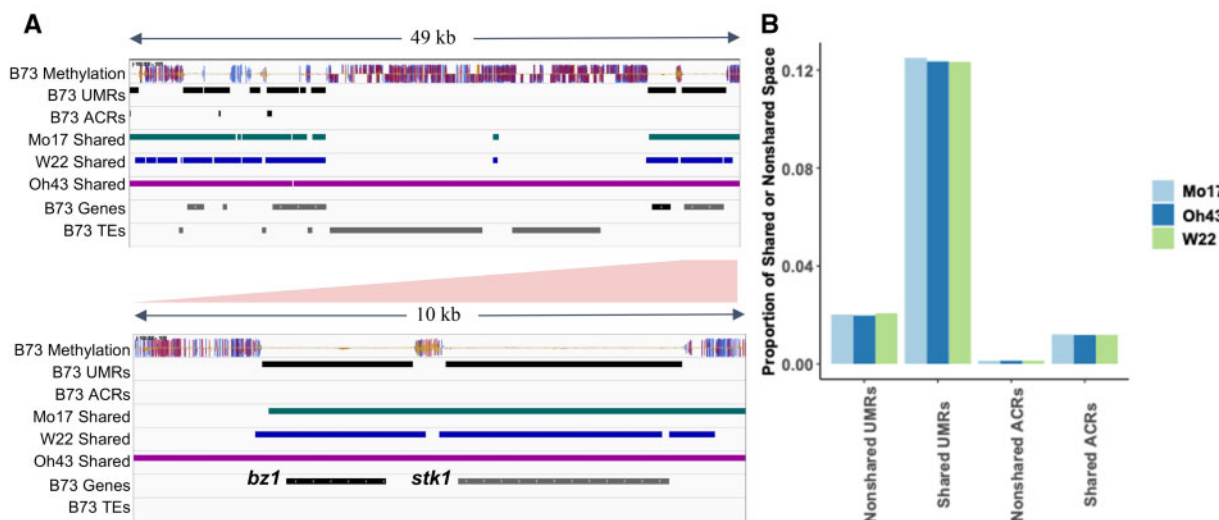
Our analysis of DNA methylation or chromatin accessibility is often focused on 100 bp bins. To directly compare the same coordinate space between genomes, we identified the 100 bp bins from the B73 genome that were shared across genotypes (Figure 2A and Supplementary Figure S4). In the comparisons of B73 to the other three genomes, we find 41–48% of the B73 bins are non-shared, 37–42% of bins have an exact match in shared regions, 12–14% mapped with  $\leq 1$  SNP, and an additional 4% mapped with  $\geq 1$  small (<20bp) indel between the two genotypes. Across all comparisons, there are over 800,000 100 bp bins that are shared in all four genotypes (Figure 2C). There are ~500,000 bins that are found only in B73 and another ~800,000 that are present in B73 and only one or two of the other two genotypes (Figure 2C). The regions that are shared between genotypes have fewer bins with missing data such that only 6.7% of the bins shared in all three genotypes lack DNA methylation data compared to 28.4% of the bins that are only present in B73. This likely reflects the fact that much of the nonshared sequence between genomes is highly repetitive and recalcitrant to unique mapping. The identification of these shared bins allowed us to calculate the methylation levels or ATAC-seq read depth for the specific coordinates in a second genome that correspond to the B73 bins to allow direct comparisons of chromatin properties between genomes using epigenomic data aligned to its own reference genome.

### UMRs and ACRs are depleted in nonshared portions of the genome

We initially focused on the chromatin properties of the non-shared portions of the genome to assess the frequency of UMRs or ACRs within the dispensable portion of the genome compared to the shared portions. While over 10% of the shared genomic regions are annotated as genic less than 4.8% of the nonshared regions are annotated as genic reflecting a depletion of genes and enrichment of intergenic and TE sequence. The analysis of the *brnze1* (*bz1*) locus on chromosome 9 illustrates these trends of shared space in genic regions and large nonshared blocks between genes, as previously described (Fu and Dooner 2002; Wang and Dooner 2006) (Figure 3A). In the *bz1* region, very few UMRs or ACRs are found within the nonshared regions (Figure 3A). We proceeded to perform a genome-wide assessment of the proportion of UMRs within shared and nonshared regions of the genome. While UMRs account for 6% of the entire B73 genome, only ~2% of the nonshared genomic regions are classified as UMRs compared to ~12% of the shared genomic regions, representing a sixfold enrichment of UMRs in shared genomic regions (Figure 3B). A similar analysis of the genome-wide distribution of ACRs reveals that accessible chromatin is 12-fold enriched within genomic regions that are shared among all four genotypes relative to nonshared regions (Figure 3B). ACRs account for 1.2% of the shared genomic space but only 0.1% of the nonshared genomic regions (Figure 3B). Both ACRs and UMRs are frequently found near genes and the nonshared genomic regions are relatively gene-poor. However, this depletion of genes is not the only explanation for the paucity of ACRs and UMRs in the nonshared genomic regions. Over 80% of the genes in the shared space contain a UMR, while only 17% of the genes located in nonshared regions



**Figure 2** Defining shared and nonshared regions between genome assemblies. (A) Schematic representation of B73-based 100bp bins defined as shared or nonshared in Mo17 and W22 (gray shaded regions) based on chromosomal alignments. The 100bp bins in W22 or Mo17 could be defined by 100bp increments within that genome sequence or based on coordinate matches to the B73 genome and these are shown as the W22 (blue) or Mo17 (purple) coordinate bins or the B73-based coordinates (grey). The black hash or the light to dark color change indicates the 100bp bin boundaries. (B) The proportion of the B73 genome that is defined as shared or nonshared with Mo17, W22, and Oh43 based on chromosome-level sequence alignments. (C) The number of B73 100bp bins that are unique to B73 (0 shared genotypes), shared with one other genotype assessed (1), shared with two other genotypes assessed (2) or shared across all 4 genotypes including B73, Mo17, Oh43, and W22 (3). Genotype labels correspond to the genotypes which share 100bp bins with B73.



**Figure 3** Presence of ACRs and UMRs within shared and nonshared genomic regions. (A) An IGV (Robinson et al. 2011) representation of a 49kb segment on chromosome 9 (upper panel) of the B73 genome assembly. Tracks show B73 methylation levels in all contexts (CG-blue, CHG-red, and CHH-yellow), B73 UMRs and ACRs, Mo17 shared sequence (green), W22 shared sequence (blue), Oh43 shared sequence (purple), and B73 gene and TE annotations (grey). The lower panel shows a closer view of a 10kb region of the *bz1* locus to see the detail. (B) The B73 genome was compared to Mo17, Oh43, or W22 to define regions that are shared or nonshared in each contrast. The proportion of the shared or nonshared space that is classified as UMR or ACR was determined for each of the pairwise contrasts.

contain UMRs. Prior studies have found that nonshared genes are less likely to be expressed (Hirsch et al. 2016; Sun et al. 2018; Anderson et al. 2019; Haber et al. 2020) and the depletion of UMRs within or near these genes further suggests that many of these features that are annotated as “genes” lack the chromatin

properties (UMRs and ACRs) that are often associated with expression. These analyses suggest that pan-genome assessment of UMRs and ACRs will provide a more complete identification of UMRs/ACRs but that there are a limited number of novel UMRs or ACRs in nonshared space in maize. The subsequent analysis



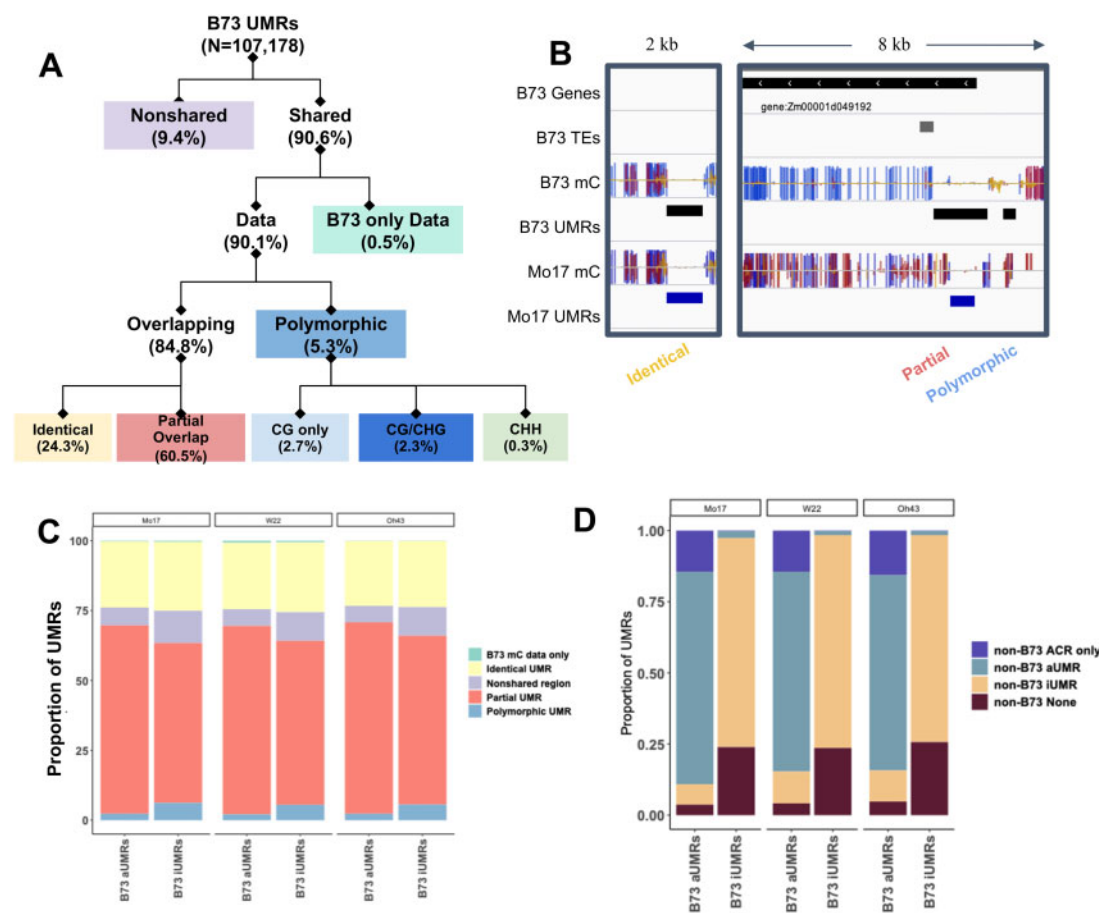
will focus on the UMRs and ACRs that are present within shared regions of any two maize genomes.

## Comparisons of UMRs and ACRs in the shared space of maize genomes

We proceeded to focus on the UMRs and ACRs that are present within shared regions between maize genomes. The analyses were primarily focused on UMRs as these encompass the vast majority of ACRs (Figure 1D) and we could monitor stability for the UMRs with an ACR (aUMRs) compared to the UMRs without an ACR (iUMRs). The B73 UMRs were compared to each of the other genomes and classified based on whether they are present in shared/nonshared regions and then whether the region has DNA methylation data available for both genotypes. For the ~90% of B73 UMRs that have defined methylation states and are present in a shared region, we could classify whether there is an overlapping UMR in the other genotype or whether the UMR is polymorphic such that it is classified as methylated in the other genotype (Figure 4A). Most UMRs that are present in shared space overlap a UMR in the other genotype while a small set (5.9%) are

polymorphic (Figure 4A). The overlapping UMRs can be classified as identical if the boundaries of the UMR are the same in both genotypes (example in Figure 4B). Alternatively, an overlapping UMR could represent a partial overlap such that one genotype has a larger region than the other or both edges are shifted (examples in Figure 4B). The UMRs with partial overlap account for the majority (71.3%) of the overlapping UMRs between two genotypes (Figure 4A). We also assessed the stability of UMR classifications between the two biological replicates of B73 data (Supplementary Figure S4). For the regions that have data in both replicates, we found very few (0.04%) polymorphic UMRs and only 0.6% of the UMRs had partial overlap between the replicates (Supplementary Figure S4).

B73 UMRs can be subdivided into aUMRs ( $n=16,627$ ) and iUMRs ( $n=91,607$ ) based on the presence, or absence, of an ACR within the UMR. We compared the distribution of classifications for the aUMRs and iUMRs for the presence of identical, partially overlapping, or polymorphic UMRs in the other genotypes (Figure 4C). The B73 aUMRs have fewer examples of polymorphic UMRs as well as fewer examples within nonshared genomic regions.



**Figure 4** Stability of UMRs in shared sequence. (A) A flowchart on how B73 UMRs are classified is shown. The numbers in parenthesis indicate the average number of regions classified in that group based on comparisons to the other genotypes. The proportion of B73 UMRs that are shared or nonshared (purple) based on sequence with the respective genome assembly. Shared regions are further classified as B73-only (green) for UMRs that lack data in the other genotype, identical (yellow) for UMRs that maintain an unmethylated state in the same region, partially overlapping (pink) for UMRs that maintain an unmethylated state but have different UMR boundaries across genotypes or polymorphic (blue) for UMRs that change to a methylated state in the other genome. The colors in A are identical to those in C. (B) A genome browser view of the several regions in the B73 genome to illustrate examples of identical, partially overlapping and polymorphic UMRs. A track of DNA methylation in all contexts (CG-blue, CHG-red, CHH-yellow) is shown for B73 and Mo17 (both aligned to B73v4) with UMRs defined below in black (B73) and blue (Mo17). B73 UMRs are defined as identical (yellow), partial overlap (pink), or polymorphic (blue). (C) The proportion of B73 UMRs that are classified in each group defined in A are shown for both aUMRs and iUMRs based on comparison to each of the other three genotypes. (D) The proportion of B73 aUMRs or iUMRs that are classified as ACR only (not unmethylated) in the other genotype (purple), aUMR in the other genotype (blue), iUMR in the other genotype (yellow), or methylated and inaccessible in the other genotype (burgundy) are shown for comparisons to each of the other genotypes



However, this is largely due to a higher proportion of overlapping UMRs that are partially overlapping rather than more examples of identical UMRs (Figure 4C). These analyses suggest that while any two genomes often have UMRs in similar regions the exact coordinates of the UMRs are often distinct.

The B73 aUMRs and iUMRs were also assessed for the potential changes to either methylation or accessibility between genotypes (Figure 4D). The majority (~71.2%) of the B73 aUMRs were maintained as aUMRs in the other genotypes. However, there are also a subset of the B73 aUMRs that lose either the unmethylated state (~14.8%) or chromatin accessibility (~11.1%) in the other genotype. The remaining 2.9% are not classified as either ACR or UMR for the same region in the other genome. The B73 iUMRs often (~73.7%) are unmethylated and inaccessible in the other genotypes (Figure 4D). There are also many (~24.5%) examples of B73 iUMRs that are methylated in the other genotype. The proportion of shifts from unmethylated to methylated states are much higher for the iUMRs than the aUMRs. Very few (~1.6%) of the B73 iUMRs exhibit accessibility in the other genotypes (Figure 4D).

### Unique properties of regions with methylation changes in various methylation contexts

While the polymorphic B73 UMRs that are methylated in another genotype only account for a small set of all UMRs (5.3%) these may represent important functional differences between genotypes. The polymorphic UMRs can be subdivided based on the prominent class of methylation in the other genotype (Figures 4A and 5A). Each of these classes of methylation changes likely reflect distinct mechanisms and chromatin types. The types of methylation observed in these regions do not reflect the genome-wide proportions of methylation types (Supplementary Figure S1). The proportions that are classified as CG only (~50%) or CHH (5–7%) are higher than observed genome wide (10 and 0.5–1% respectively) (Supplementary Figures S1 and S5B). The remaining ~40% of the polymorphic UMRs exhibit CG/CHG methylation in the genotype that has methylation (Figure 5B).

The presence or absence of ACRs in both genotypes was assessed for the polymorphic UMRs relative to overlapping UMRs (Figure 5C). While both identical UMRs and partially overlapping UMRs show virtually identical proportions with shared ACRs or polymorphic ACRs, the polymorphic UMRs have very few stable ACRs (Figure 5C). This is expected as there are very few examples of accessible regions within methylated DNA. The proportion of the polymorphic UMRs that are classified as having an ACR only in B73 but not the methylated genotype is quite variable. Polymorphic UMRs with CHH methylation in the other genotype are more likely to have an ACR in B73 than polymorphic UMRs with CG only methylation in the other genotype (Figure 5C). This could reflect the fact that CHH methylation is often found in regions immediately upstream or downstream of genes in the maize genome (Gent et al. 2013; Li et al. 2015a) and that these regions often have ACRs. In contrast, the CG-only methylation often occurs within gene bodies, where ACRs are less common than at the edges of genes or promoter regions.

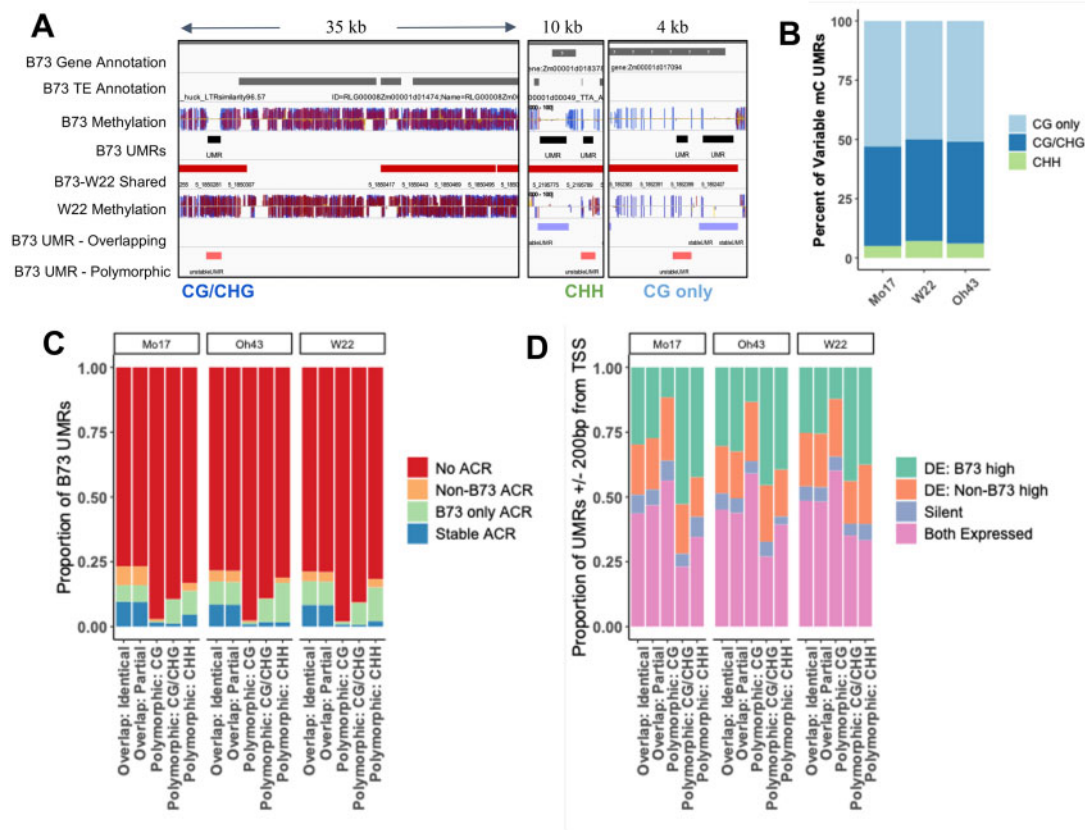
We proceeded to assess variable gene expression of genes near overlapping or polymorphic UMRs using RNA-seq data from the same tissue used to monitor accessibility and DNA methylation. Genes with an overlapping or polymorphic UMR within 200bp (upstream or downstream) of the transcription start site (TSS) were identified and classified as being differentially expressed (DE), expressed in both genotypes but not DE, or silent (FPKM < 1 in both genotypes). The sets of genes that have identical or partially overlapping UMRs near the TSS exhibit nearly identical

proportions of genes in these categories suggesting little functional difference between identical and partially overlapping UMRs (Figure 5D). Polymorphic UMRs that gain CG-only methylation in the other genotype have fewer examples (~12%) of genes with higher expression in B73 compared to identical or partially overlapping UMRs (25–32%). More of these polymorphic UMRs with CG-only methylation exhibit expression in both genotypes (56–60%) compared to the UMRs with identical overlap (43–48%). The percent of genes that are more highly expressed in B73 (which is unmethylated) than in the other genotype (which has methylation) is higher for genes with gains of CG/CHG (44–52%) or CHH (37–42%) methylation compared to the genes with identical UMRs (25–29%) suggesting that a subset of these methylation gains may be associated with reduced expression. While the presence of polymorphic CHH or CG/CHG methylation near the TSS has an enrichment for genes that are more highly expressed in the unmethylated genotype there are still 42–57% of these genes that have an equivalent expression in the two genotypes or higher expression in the genotype with higher methylation. This suggests that the gain of CG/CHG methylation or CHH methylation in regions surrounding the TSS can be associated with altered expression in some cases, but that other genes can tolerate variable methylation without a significant change in expression.

### Partially overlapping UMRs contribute substantially to differentially methylated regions

The analysis of natural variation for DNA methylation is often focused on identification of differentially methylated regions (DMRs) between genotypes. In this study, we elected to focus on the conservation/variation of UMRs as these regions have evidence for functional relevance in crop genomes. However, the observation that many of these regions only have partial overlap suggests that many DMRs might be the result of a shift in the boundary between methylated and unmethylated DNA rather than a complete regional gain/loss of methylation (Figure 6A). The 100bp bins were used to identify DMRs between the genotypes. There are 116,000–158,000 100bp bins that are classified as differentially methylated with hypomethylation in B73 relative to the other genotype. We assessed how many of these DMRs are due to completely polymorphic UMRs compared to partial UMRs with different boundaries between methylated and unmethylated DNA (Figure 6B). The polymorphic UMRs account for 2.5–3.3% of all differentially methylated bins depending on which genotypes are being compared. A larger proportion (51.5–53.5%) of the differentially methylated bins are due to partially overlapping UMRs. The remaining differentially methylated bins occur in regions too small to be classified as UMRs (UMRs <300bp) or represent single bin differences in larger UMRs. This analysis suggests that many of the DMRs are due to shifting boundaries between methylated and unmethylated DNA rather than a complete gain or loss of methylation in a region. It is noteworthy that within a genotype we find very few examples of shifting boundaries between biological replicates (Supplementary Figure S4).

These observations suggest that the specific boundary between methylated and unmethylated DNA can be variable between genotypes. This could be due to sequence changes at or near the edges of these regions or could arise due to stochastic variation with no sequence change. To address this question we assessed the proportion of identical or partially overlapping UMRs within large (>1Mb) blocks of sequence that is IBS. In total there was 112.7Mb of IBS sequence blocks that could be assessed and these are large blocks of sequence that are essentially devoid of SNPs or structural variants. While 5.3% of all UMRs are classified as polymorphic we



**Figure 5** Characteristics of polymorphic UMRs. All B73 UMRs classified as polymorphic (shown in Figure 4A) were assessed based on the type of methylation present in the methylated genotype. The classification is based on which type of methylation state is most common within the 100 bp bins of the UMR. (A) A genome browser view of a region on chromosome 5 of the B73 genome. A track of B73 methylation in all contexts (CG-blue, CHG-red, CHH-yellow) is shown with UMRs defined below in black. Regions with shared sequence with W22 are shown in red and the W22 methylation track (aligned to the B73v4 assembly) with corresponding UMR classification as overlapping (purple) or polymorphic (red). Three separate snapshots are shown with the type of methylation found in W22 for the variable UMR noted below (CG only, CG/CHG, or CHH). (B) The percent of all B73 UMRs classified as polymorphic that change to CG only (light blue), CG/CHG (dark blue), or CHH (green) methylation in the other genotype was calculated. (C) UMRs were defined as containing an ACR in both genotypes (Stable ACR: blue), in one genotype (B73 only ACR: green, NonB73 ACR: orange), or lacking an ACR in both genotypes (No ACR: red). The proportion of each category of B73 UMR (overlapping and polymorphic) that is defined by ACR presence or absence is shown for each genotype. (D) The proportion of UMRs that are found within 200 bp of an annotated gene TSS that are defined as differentially expressed (DE), expressed in both genotypes or not expressed is shown for each genotype. Genes were classified as differentially expressed ( $\log_2$  fold change  $> 2$  and  $P$ -value  $< 0.05$ ) with the higher expression level observed in B73 (green) or the nonB73 genotype (orange) or as nondifferentially expressed (FPKM  $> 1$ , pink) or not expressed (silent: purple).

find only 2.8% of UMRs that are classified as polymorphic in these regions suggesting that fully polymorphic UMRs are depleted (1.9-fold) in the absence of sequence variation (Figure 6C). The IBS regions have a higher proportion of UMRs with identical boundaries in the two genotypes. However, there are still a large number of UMRs with shifted boundaries (49.5%) suggesting that the boundaries between methylated and unmethylated DNA can shift even without nearby sequence variation.

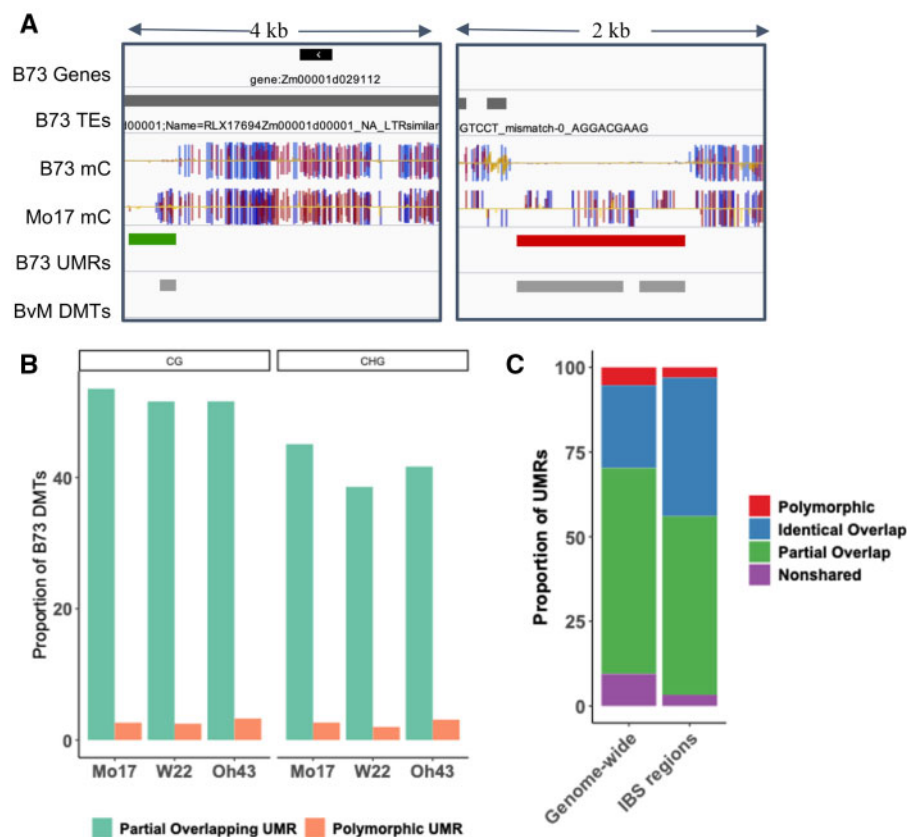
## Discussion

*Zea mays*, unlike many other model organisms, has a large genome containing 80% repetitive sequence and high levels of DNA methylation interspersed with functional genic and regulatory regions (Schnable et al. 2009; Jiao et al. 2017). Examination of genome structure across inbred lines have identified extensive polymorphism in both genic and repeat regions of the maize genome (Chia et al. 2012; Hirsch et al. 2014; Springer et al. 2016; Darracq et al. 2018; Anderson et al. 2019; Hufford et al. 2021). Prior analyses of natural variation of chromatin in maize have been based on epigenome profiling data aligned to a single reference genome (Li et al. 2015b;

Xu et al. 2020). While a single reference genome provides insight into variation in conserved genomic regions, it does not contain the full set of sequences present in the lines being compared, resulting in biases in the ability to compare chromatin properties. The availability of multiple *de novo* genome assemblies allows for a more complete discovery of regions with specific chromatin properties, such as UMRs or ACRs. In this study, we profiled genome-wide DNA methylation, based on alignments of data to the corresponding genome assembly, to identify the ~6% of each genome that exhibits an unmethylated state and the ~1% that is accessible chromatin. A pan-genomic analysis of UMRs and ACRs reveals the frequency of these features within both shared and nonshared genomic regions. Within the shared sequence regions it is possible to assess the stability of the unmethylated and accessible chromatin portions of the genome.

## Pan-genome analyses reveal enrichment of unmethylated regions within shared sequence

In a comparison of any two genomes, the sequence unique to each genome is primarily composed of highly repetitive sequences with extensive DNA methylation and is found to be depleted



**Figure 6** Many DMTs are due to partially overlapping UMRs. (A) IGV (Robinson et al., 2011) view of DMTs. Tracks show B73 gene and TE annotations, B73 and Mo17 single cytosine methylation in all contexts (CG: blue, CHG: red, CHH: yellow), B73 UMRs and classification relative to Mo17 (identical: blue, partial: green, polymorphic: red), and DMTs defined by a low level of B73 CG methylation and high level of Mo17 CG methylation. (B) The proportion of B73 DMTs that are associated with partially overlapping UMRs (green) or polymorphic UMRs (orange) is shown. (C) The proportion of B73 UMRs, genome-wide (control) or in IBS regions, that are shared or nonshared (purple) based on sequence with the respective genome assembly. Shared regions are further classified as missing data (orange) for UMRs that lack data in the other genome, identical (blue) for UMRs that maintain an unmethylated state in the same region, partially overlapping (green) for UMRs that maintain an unmethylated state but have different UMR boundaries across genotypes or polymorphic (red) for UMRs that change to a methylated state in the other genome.

for genes (Chia et al. 2012; Hirsch et al. 2016; Springer et al. 2016; Darracq et al. 2018; Anderson et al. 2019; Hufford et al. 2021). The proportion of the nonshared genome that is classified as UMR or ACR is 6–12 fold lower than the proportion of the shared genome classified as UMR or ACR. This reduction in UMRs and ACRs is not simply due to the reduced gene content in nonshared space. Most (80%) of genes in the shared space are associated with a UMR, while only 17% of genes in nonshared space have a UMR. This is not unexpected as prior studies of presence-absence variation (PAV) genes have found that most of these genes that vary between genotypes are not expressed even when they are present (Swanson-Wagner et al. 2010). A recent analysis of 26 maize genomes that used a slightly different approach to classify unmethylated and CG-only regions reported similar findings (Hufford et al. 2021). More UMRs are identified by aligning chromatin data to the proper genome but the proportion of UMRs or ACRs in this nonshared space is much lower than in the shared regions. These analyses suggest that pan-genomic analyses can identify novel UMRs or ACRs but that these are relatively rare in the sequences that exhibit large scale structural variation. However, it is worth noting that the UMRs or ACRs that are present near the genes in the nonshared space can be an effective tool for identifying genes with potential expression (Sartor et al. 2019; Crisp et al. 2020). Given that many of the genes within these regions are likely pseudogenes generated by transposition of genes or gene

fragments that can be difficult to annotate just based on sequence, the use of chromatin data can help to identify genes with potential function in these regions.

### Characterization of relative dynamics of accessibility and methylation

We were interested in studying the relative dynamics of both DNA methylation and chromatin accessibility among genotypes. Prior studies have found that the majority of accessible regions have little or no methylation (Ricci et al. 2019) but that there are also many UMRs that lack accessibility (Crisp et al. 2020). The analysis of UMRs that are present within shared sequence regions can be used to understand how often there is variation in only accessibility as opposed to coordinate changes in both DNA methylation and accessibility. The accessible UMRs (aUMRs) tend to be relatively stable in other genotypes with both accessibility and lack of DNA methylation for an overlapping region in other haplotypes. This is consistent with the concept that these regions may be important for proper regulation of gene expression and therefore changes in these chromatin properties could be associated with functional differences. The inaccessible UMRs (iUMRs) were often inaccessible and unmethylated in both genotypes but there were a large number of these that exhibit polymorphic DNA methylation status such that they exhibit high levels of DNA methylation in the other genotypes. Only a small

proportion of these UMRs exhibit a consistently unmethylated state in both genotypes with accessible chromatin in only one of the two genotypes. These likely include some examples of false negatives due to relatively stringent criteria for calling an ACR. In these cases, an ACR may be present in both genotypes but only identified as significant for one genotype. However, these cases of variable chromatin accessibility also include examples with clear support for chromatin accessibility in one genotype but no evidence for chromatin accessibility in the other genotype. These are interesting as they potentially reflect differences in transcription factor occupancy for regions that are stably unmethylated in both genotypes. It is possible that these may reflect differences in tissue-specific expression patterns of some maize genes. In leaf tissue, there may be differential chromatin accessibility, but it is possible that the genotype without chromatin accessibility in leaf tissue still becomes accessible in some other tissue that exhibits expression. Alternatively, minor sequence changes at transcription factor binding sites may result in loss of chromatin accessibility even though the region is unmethylated in both genotypes.

### Stability and instability of UMRs between genotypes

A subset of the shared sequence UMRs do not maintain their unmethylated state across genotypes and instead have high levels of methylation in at least one of the other three genotypes. The presence of methylation variation in the shared sequence regions allowed for characterization of attributes associated with chromatin state instability. Prior studies have suggested that structural variants, especially transposable element polymorphisms, can be associated with changes in DNA methylation for nearby sequences (Eichten *et al.* 2012; Schmitz *et al.* 2013). When analyzing DNA methylation based on a single reference genome it can be difficult to incorporate information about structural variants and to map reads near the junctions of these variants. Using alignments to each reference genome and then comparing coordinates of syntenic 100 bp tiles allowed us to monitor changes in DNA methylation between genotypes, even in regions near structural variants. The polymorphic UMRs that represent a full shift of an UMR in one genotype to methylation in the other genotype are depleted in regions devoid of structural variants. Within large blocks of IBS 2.8% of the UMRs are polymorphic. In contrast, over 5.3% of all UMRs are classified as polymorphic. This indicates that changes in methylation state can occur in the absence of nearby structural variants but that the rate is substantially higher in regions with sequence variation.

In this study, we focused on the conservation and variation for UMRs or ACRs between genotypes. These are relatively large (at least 300 bp based on the criteria used for discovery) regions that lack DNA methylation. We focused on these regions due to prior evidence for functional enrichment of these regions (Oka *et al.* 2017; Ricci *et al.* 2019). We note that most of the UMRs in one genotype have an overlapping UMR in another genotype. This suggested stability of these chromatin patterns among genotypes. However, closer inspection revealed that the majority of these overlapping UMRs have different boundaries in the two genotypes. These include examples in which one UMR is entirely within the other as well as examples that have partial overhangs in both genotypes. The partially overlapping UMRs seem to have very similar genomic distributions and overlap with ACRs or altered gene expression in similar proportion to those for identical conserved UMRs. This suggests that these shifts in the boundary between methylated and unmethylated DNA do not have functional impact in most cases. This may suggest that the presence

of a UMR is more defined by sequences in the middle of the UMRs rather than particular sequences at the edges that define the extent of methylation.

The observation of many partially overlapping UMRs suggested that these shifts in the boundary between methylated and unmethylated DNA could account for many examples of differential methylation between genotypes. Conceptually, it is tempting to think that most differentially methylated regions result from a local gain or loss of a patch of DNA methylation. However, our analyses suggest that many of the differentially methylated 100 bp tiles actually arise due to changes in the boundaries between UMRs in different genotypes. Further studies will be necessary to determine if these differences in methylation boundaries represent a continuum such that each genotype has a slightly different boundary or if there are preferred epi-haplotypes.

### Acknowledgments

The authors would like to acknowledge Christina Ethridge for preparation of ATAC-seq libraries and Pete Hermanson for preparation of WGBS libraries. They acknowledge the Minnesota Supercomputing Institute (MSI) at the University of Minnesota for providing computational resources that contributed to the research results reported within this paper.

### Funding

This study was funded by support from the National Science Foundation IOS-1802848 to N.M.S. and R.J.S., as well as IOS-1856627 to R.J.S. A.P.M. was supported by an NSF Postdoctoral Fellowship in Biology (DBI-1905869).

### Conflicts of interest

None declared.

### Literature cited

- Anders S, Pyl PT, Huber W. 2015. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*. 31: 166–169.
- Anderson SN, Stitzer MC, Brohammer AB, Zhou P, Noshay JM, *et al.* 2019. Transposable Elements Contribute to Dynamic Genome Content in Maize. *Plant Journal*. 100:1052–1065.
- Anderson SN, Zynda G, Song J, Han Z, Vaughn M, *et al.* 2018. Subtle perturbations of the maize methylome reveal genes and transposons silenced by Chromomethylase or RNA-directed DNA Methylation Pathways. *G3 (Bethesda)*. 8:1921–1932.
- 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.
- Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods*. 10:1213–1218.
- Chia JM, Song C, Bradbury PJ, Costich D, de Leon N, *et al.* 2012. Maize HapMap2 identifies extant variation from a genome in flux. *Nat Genet*. 44:803–807.
- Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, *et al.* 2008. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*. 452:215–219.



- Crisp PA, Marand AP, Noshay JM, Zhou P, Lu Z, et al. 2020. Stable unmethylated DNA demarcates expressed genes and their cis-regulatory space in plant genomes. *Proc Natl Acad Sci USA*. 117:23991–24000.
- Darracq A, Vitte C, Nicolas S, Duarte J, Pichon J-P, et al. 2018. Sequence analysis of European maize inbred line F2 provides new insights into molecular and chromosomal characteristics of presence/absence variants. *BMC Genomics*. 19:119.
- Eichten SR, Briskine R, Song J, Li Q, Swanson-Wagner R, et al. 2013. Epigenetic and genetic influences on DNA methylation variation in maize populations. *Plant Cell*. 25:2783–2797.
- 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.
- Eichten SR, Swanson-Wagner RA, Schnable JC, Waters AJ, Hermanson PJ, et al. 2011. Heritable epigenetic variation among maize inbreds. *PLoS Genet*. 7:e1002372.
- Fu H, Dooner HK. 2002. Intraspecific violation of genetic colinearity and its implications in maize. *Proc Natl Acad Sci USA*. 99: 9573–9578.
- Gent JI, Ellis NA, Guo L, Harkess AE, Yao Y, et al. 2013. CHH islands: *de novo* DNA methylation in near-gene chromatin regulation in maize. *Genome Res*. 23:628–637.
- Haberer G, Kamal N, Bauer E, Gundlach H, Fischer I, et al. 2020. European maize genomes highlight intraspecies variation in repeat and gene content. *Nat Genet*. 52:950–957.
- 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 CN, 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*. 28: 2700–2714.
- Hoefsloot HC, Stam ME. 2020. In plants distal regulatory sequences overlap with unmethylated rather than low-methylated regions, in contrast to mammals. *bioRxiv*.
- Hufford MB, Seetharam AS, Woodhouse MR. 2021. *De novo* assembly, annotation, and comparative analysis of 26 diverse maize genomes. *bioRxiv*.
- Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, et al. 2017. Improved maize reference genome with single-molecule technologies. *Nature*. 546:524–527.
- Kawakatsu T, Stuart T, Valdes M, Breakfield N, Schmitz RJ, et al. 2016. Unique cell-type-specific patterns of DNA methylation in the root meristem. *Nat Plants*. 2:16058.
- Kent WJ. 2002. BLAT—the BLAST-like alignment tool. *Genome Res*. 12:656–664.
- Kim D, Langmead B, Salzberg SL. 2015. *hisat2*. *Nat Methods*. 12: 357–360.
- Li H. 2018. *Minimap2*: pairwise alignment for nucleotide sequences. *Bioinformatics*. 34:3094–3100.
- Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 27:2987–2993.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. 2009. The Sequence alignment/map format and SAMtools. *Bioinformatics*. 25:2078–2079.
- Li Q, Gent JI, Zynda G, Song J, Makarevitch I, et al. 2015a. RNA-directed DNA methylation enforces boundaries between heterochromatin and euchromatin in the maize genome. *Proc Natl Acad Sci USA*. 112:14728–14733.
- Li Q, Song J, West PT, Zynda G, Eichten SR, et al. 2015b. Examining the causes and consequences of context-specific differential DNA methylation in maize. *Plant Physiol*. 168:1262–1274.
- Liang Z, Anderson SN, Noshay JM, Crisp PA, Enders TA, et al. 2021. Genetic and epigenetic variation in transposable element expression responses to abiotic stress in maize. *Plant Physiology*. 186: 420–433.
- Lister R, Ecker JR. 2009. Finding the fifth base: genome-wide sequencing of cytosine methylation. *Genome Res*. 19:959–966.
- Love M, Anders S, Huber W. 2014. Differential analysis of count data—the DESeq2 package. *Genome Biol*. 15:10–1186.
- Lu Z, Hofmeister BT, Vollmers C, DuBois RM, Schmitz RJ. 2017. Combining ATAC-seq with nuclei sorting for discovery of cis-regulatory regions in plant genomes. *Nucleic Acids Res*. 45:e41.
- Marand AP, Chen Z, Gallavotti A, Schmitz RJ. 2020. A cis-regulatory atlas in maize at single-cell resolution. *bioRxiv*. 10.1101/2020.09.27.315499
- Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *Embnet J*. 17:10–12.
- Niederhuth CE, Bewick AJ, Ji L, Alabady MS, Kim KD, et al. 2016. Widespread natural variation of DNA methylation within angiosperms. *Genome Biol*. 17:194 10.1186/s13059-016-1059-0PMC: 27671052
- 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.
- 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.
- Regulski M, Lu Z, Kendall J, Donoghue MT, Reinders J, et al. 2013. The maize methylome influences mRNA splice sites and reveals widespread paramutation-like switches guided by small RNA. *Genome Res*. 23:1651–1662.
- 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.
- Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, et al. 2011. Integrative genomics viewer. *Nat Biotechnol*. 29:24–26.
- Rodgers-Melnick E, Vera DL, Bass HW, Buckler ES. 2016. Open chromatin reveals the functional maize genome. *Proc Natl Acad Sci USA*. 113:E3177–E3184.
- Sartor RC, Noshay J, Springer NM, Briggs SP. 2019. Identification of the expressome by machine learning on omics data. *Proc Natl Acad Sci USA*. 116:18119–18125.
- Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, et al. 2013. Patterns of population epigenomic diversity. *Nature*. 495: 193–198.
- Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science*. 326:1112–1115.
- 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*.
- 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*. 18:563–575.
- Springer NM, Ying K, Fu Y, Ji T, Yeh CT, et al. 2009. Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet*. 5: e1000734.
- Stitzer MC, Anderson SN, Springer NM, Ross-Ibarra J. The Genomic Ecosystem of Transposable Elements in Maize.

- 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.
- Swanson-Wagner RA, Eichten SR, Kumari S, Tiffin P, Stein JC, et al. 2010. Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Res.* 20:1689–1699.
- The Arabidopsis Genome Initiative 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature.* 408: 796–815.
- 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.
- West PT, Li Q, Ji L, Eichten SR, Song J, et al. 2014. Genomic distribution of H3K9me2 and DNA methylation in a maize genome. *PLoS One.* 9:e105267.
- Xi Y, Li W. 2009. BSMAP: whole genome bisulfite sequence MAPPING program. *BMC Bioinformatics.* 10:232.
- Xu J, Chen G, Hermanson PJ, Xu Q, Sun C, et al. 2019. Population-level analysis reveals the widespread occurrence and phenotypic consequence of DNA methylation variation not tagged by genetic variation in maize. *Genome Biol.* 20:243.
- Xu G, Lyu J, Li Q, Liu H, Wang D, et al. 2020. Evolutionary and functional genomics of DNA methylation in maize domestication and improvement. *Nat Commun.* 11:5539.

Communicating editor: E. Akhunov