G3, 2021, 11(10), jkab263
DOI: 10.1093/g3journal/jkab263
Advance Access Publication Date: 27 July 2021
Genome Report
Featured

Leveraging histone modifications to improve genome annotations

John Pablo Mendieta,[1] Alexandre P. Marand ,[1] William A. Ricci,[2] Xuan Zhang,[1] and Robert J. Schmitz [1,*]

[1]Department of Genetics, University of Georgia, Athens, GA 30602, USA and
[2]Department of Plant Biology, University of Georgia, Athens, GA 30602, USA

*Corresponding author: Email: schmitz@uga.edu

Abstract

Accurate genome annotations are essential to modern biology; however, they remain challenging to produce. Variation in gene structure and expression across species, as well as within an organism, make correctly annotating genes arduous; an issue exacerbated by pitfalls in current *in silico* methods. These issues necessitate complementary approaches to add additional confidence and rectify potential misannotations. Integration of epigenomic data into genome annotation is one such approach. In this study, we utilized sets of histone modification data, which are precisely distributed at either gene bodies or promoters to evaluate the annotation of the *Zea mays* genome. We leveraged these data genome wide, allowing for identification of annotations discordant with empirical data. In total, 13,159 annotation discrepancies were found in *Z. mays* upon integrating data across three different tissues, which were corroborated using RNA-based approaches. Upon correction, genes were extended by an average of 2128 base pairs, and we identified 2529 novel genes. Application of this method to five additional plant genomes identified a series of misannotations, as well as identified novel genes, including 13,836 in *Asparagus officinalis*, 2724 in *Setaria viridis*, 2446 in *Sorghum bicolor*, 8631 in *Glycine max*, and 2585 in *Phaseolous vulgaris*. This study demonstrates that histone modification data can be leveraged to rapidly improve current genome annotations across diverse plant lineages.

Keywords: epigenomics; genome annotation; histone modification; maize; plant genomes

Introduction

Accurate genome annotations and assemblies are an essential resource for modern biology. Their capacity to facilitate genetic inquiry, as well as operate as the backbone for genome biology makes their production vital. However, while the creation of gapless, and near-perfect genome assemblies is becoming commonplace (Liu *et al.* 2020; Miga *et al.* 2020), genome annotation remains challenging (Salzberg 2019). Generation of a genome annotation requires multiple lines of evidence in the form of mRNA expression data, homology-based inference, and *in silico* prediction algorithms, which are synthesized into a single concordant annotation (Yandell and Ence 2012). The challenges of such complex data synthesis, potentially compounded by the generation of *in silico* artifacts at each aforementioned stage of analysis, make accurate genome annotation precarious at best (Salzberg 2019).

The epigenome provides an invaluable untapped resource which adds additional support to increase confidence in genome annotation. Generally, eukaryotic genomes are divided into two distinct domains, (1) euchromatin, which is gene-rich and has abundant transcriptional activity, and (2) heterochromatin, which is gene-poor, densely populated with repeats and transposable elements and mostly devoid of transcriptional activity (McClintock 1950; Hannah 1951). These two major domains of the epigenome are defined by their occurrence with specific covalent modifications to DNA and to the alpha globulin tail of histones, which together comprise chromatin (Luger 1997). Histone modifications are diverse and they correlate with a wide range of biological phenomena. Some have proposed that chromatin comprises a "language" or code all its own in the genome, with different combinations and permutations of histone modifications correlating to distinct biological outputs (Strahl and Allis 2000; Rando 2012). Evolutionarily, histone modifications and their functions are deeply conserved, with eukaryotes using similar sets of histone modifications around transcriptionally active and inactive regions of the genome (Schübeler *et al.* 2004; Bernstein *et al.* 2005; Morris *et al.* 2007), suggesting their essentiality to eukaryotic genomes.

In plants specifically, recent large-scale studies have corroborated histone modification function, and co-localization to specific regions of the genome (Shi and Dawe 2006; Li *et al.* 2008; Mahrez *et al.* 2016; Lu *et al.* 2019; Ricci *et al.* 2019). For example, transcribed genes generally possess Histone H3 Lysine 4 trimethylation (H3K4me3) and Histone H3 Lysine 9/27/56 (H3K9/27/56ac) acetylation near their transcriptional start sites and H3K4me1 and H3K36me3 throughout their gene bodies (Zhang *et al.* 2009; Roudier *et al.* 2011; Li *et al.* 2015; Oka *et al.* 2017; Lu *et al.* 2019; Ricci *et al.* 2019), whereas actively silenced genes often possess Histone H3 Lysine 27 trimethylation (H3K27me3) throughout their gene bodies and promoter-proximal regions (Zhang *et al.* 2006, 2007; Zilberman *et al.* 2007; Bernatavichute *et al.* 2008;

Received: May 21, 2021. **Accepted:** July 15, 2021

Li et al. 2008). The epigenomic landscape of heterochromatin is quite distinct, as repeats and transposable elements are highly enriched for DNA methylation, H3K9me2, and small RNAs (Lu et al. 2005; Zhang et al. 2007; Zilberman et al. 2007; Bernatavichute et al. 2008). The unique patterns and distributions of histone modifications throughout the genome, especially within transcribed genes, provides a unique opportunity to improve efforts in genome annotations.

Histone modifications associated with transcription reflect various features of transcriptional units. For example, in *Arabidopsis thaliana*, H3K4 can be either mono- di- or tri-methylated by ARABIDOPSIS HOMOLOG OF TRITHORAX1 (ATX1) and ARABIDOPSIS HOMOLOG OF TRITHORAX2 (ATX2), (Nislow et al. 1997; Alvarez-Venegas et al. 2003; Saleh et al. 2008). These histone modifications primarily occur at genic regions of the genome, with H3K4me2 and H3K4me3 being distributed specifically around transcriptional start sites (Zhang et al. 2009). H3K4me3 as well as ATX1 are also found tightly linked to Pol II occupancy, as ATX1 and specific subunits of Pol II are consistently found to co-localize at promoters (Fromm and Avramova 2014). Binding of ATX1 and Pol II form a transcriptional initiation complex, allowing for rapid transcriptional responses (Song et al. 2015). Paired with this, increased proportions of H3K4me3 at promoters correlate with enhanced transcriptional rates (Zhang et al. 2009).

A histone modification which is intimately linked to transcription elongation is Histone H3 Lysine 36 methylation. During transcription the phosphorylated carboxy terminal domain of RNA Pol II recruits the histone methyltransferase Su(var)3-9, Enhancer-of-zeste and Trithorax 2, or SET2 (homolog SET DOMAIN GROUP 8, or SDG8 in *A. thaliana*), to methylate H3K36 (Wagner and Carpenter 2012). Much like H3K4, H3K36 can be mono- di- or tri-methylated, but only di- and tri-methylation correlate with transcription in plants (Xu et al. 2008). SET2 limits the occupancy of Pol II in yeast, indicating its essential role during transcription elongation (Kizer et al. 2005). In *A. thaliana*, mutation of *SDG8* has been implicated in a range of phenotypic phenomena from development, to timing of flowering (Cartagena et al. 2008; Cazzonelli et al. 2009; Bu et al. 2014; Jin et al. 2015). In plants, H3K36me3 co-occurs with the length of the transcribed units, demonstrating its deeply conserved function (Li et al. 2008; Lu et al. 2019). Uniquely in plant genomes, H3K36me3 is correlated with the histone modification H3K4me1 across the length of the transcribed unit (Zhang et al. 2009; van Dijk et al. 2010; Ricci et al. 2019). This is in stark contrast to metazoan genomes where H3K4me1 primarily denotes intergenic enhancers (Bannister and Kouzarides 2011; Rada-Iglesias et al. 2011).

Unlike the histone residues which are methylated, histone residues which are acetylated have a direct functional impact on transcription. Whereas methylated histones often act indirectly by recruiting protein complexes that impact the chromatin landscape, acetylated histones physically alter how DNA wraps around the nucleosome (Allfrey et al. 1964; He et al. 2003). The negatively charged acetyl groups added on the histone protein repel negatively charged DNA promoting a more permissive environment for transcription (Allfrey et al. 1964; Earley et al. 2007). In plant genomes, acetylated histones co-occur with other transcription initiation histone modifications, such as H3K4me3 around the promoter sequence of actively transcribed genes (Roudier et al. 2011; Lu et al. 2019). Interestingly, in plants, histone acetylation can also indicate accessible chromatin in proximal and distal cis-regulatory elements (Oka et al. 2017; Zhao et al. 2018; Lu et al. 2019; Ricci et al. 2019).

Previous studies demonstrated that histone modification data can be leveraged on a genome-wide scale for a multitude of uses. For instance, in Sartor et al. epigenomic data was used to identify the expressed regions of the maize genome, in what is sometimes called the "expressome" (Sartor et al. 2019). This utilization further allowed them to identify regions of the maize genome, which are likely functional, and not constitutively repressed. However, although this leveraging of epigenomic data provides valuable insights into expressed regions of the genome, it does not seek to amend potential annotations issues present in current annotations (Sartor et al. 2019). Histone modification data has been used to annotate regions of the genome potentially harboring unannotated genes, or long noncoding RNAs (lncRNAs) (Guttman et al. 2009; Jarroux et al. 2017). A recent analysis of the *Z. mays* epigenome identified signals of actively transcribed transcriptional units outside of currently annotated gene features (Ricci et al. 2019). However, to date, few studies have leveraged epigenomics to improve the quality of genome annotations (Dozmorov 2017; Ernst and Kellis 2017). In this study, we show that integration of RNA-sequencing (RNA-seq) data with histone modification data significantly improves genome annotations.

Methods

Genome versions and annotation

The maize genome V4 and annotation set version 4.38 of the annotation were acquired from gramene and used for all analysis (Jiao et al. 2017). The asparagus genome was taken from the asparagus genome project (http://asparagus.uga.edu/tripal/, Last accessed 8.2.2021). The genomes for other genomes were retrieved from phytozome version 13 with the most recent annotations used.

ChIP-seq data processing peaks

Raw reads from five different ChIP-seq libraries consisting of two replicates each were used to identify regions of enrichment for the histone modifications H3K36me3, H3K4me1, H3K56ac, and H3K4me3, as well input genomic. Reads were trimmed using trimmomatic, and aligned to the genome using bowtie2, "–very-sensitive" (Langmead and Salzberg 2012; Bolger et al. 2014). Only uniquely mapping reads were used for downstream analysis. Peak calling was done for histone modifications known to have broad peaks (H3K36me3 and H3K4me1) using the software epic2 with the parameters "–false-discovery-rate-cutoff .1 –keep-duplicates," as well as MACS2 to identify smaller regions of enrichment using the parameters "callpeak –keep-dup all -g 1.6e9 -q .1." Narrow peaks (H3K56ac and H3K4me3) were called using MACS2 with the parameters "–keep-dup all –extsize 147 -g 1.6e9 -p .05" (Zhang et al. 2008; Stovner and Sætrom 2019). Peaks which were within 480 bps of each other were merged. Intersection between replicates of the same histone modification were taken. The minimum and maximum distanced regions were taken between intersecting regions, and the results merged to give a single peak which overlapped the extent of both peaks.

RNA-seq data processing

Raw reads were trimmed using trimmomatic with default parameters (Bolger et al. 2014). Reads were aligned to the *Z. Mays* reference genome version 4 using the STAR aligner, and the values "–outSAMstrandField intronMotif –outSAMmapqUnique 255 –alignIntronMax 50000" (Dobin et al. 2013). TPM values were calculated using TPMCalculator from NCBI.

Generation of heatmaps and metaplots

ChIP-seq data were handled similarly to ChIP-seq peak calling, with only uniquely mapping reads used. Libraries were normalized by read number using the "bamCoverage" command found in deepTools version 3.3.1, and normalized using counts per million (CPM) mapped reads (Ramírez *et al.* 2014). Matrices were generated with the compute matrix function "scale-regions" with parameters "-bs 20 -b 1000 -a 1000 –regionBodyLength 5000." Matrices were loaded into a custom R script and the R library EnrichedHeatMap was used to plot heatmaps (Gu *et al.* 2018). Genomic input reads were subtracted from ChIP-seq signal to account for genome bias, and the 95% quantile of each data set was selected as the upper value.

Mappability control

In order to ensure that we were controlling for potential mappability issues in our analysis we utilized Genmap version 1.3.0 (Pockrandt *et al.* 2020). We generated mappability scores at single base pair resolution for unique kmer size 75 (size of our ChIP-seq reads) for the entire maize genome using the flags "-K 75."

Annotating the genome using ChIP-seq

A custom pipeline was developed to annotate the genome using peak calls from ChIP-seq. Current annotations were categorized as either being expressed, or unexpressed based on alignment of stranded RNA-seq reads (Greater than 5 RNA-seq reads), as well as overlapping peak calls correlating with gene body extensions (H3K36me3 and H3K4me1). Annotations were considered "good" or "unaltered" if histone modifications H3K36me3 or H3K4me1 overlapped the length of the gene body, and the annotation overlapped a peak correlating with promoter transcription initiation (H3K56ac or H3K4me3) in the first 50% of the gene body. Expressed annotations which did not contain a peak correlating with a promoter were then further explored by searching upstream of the transcription start site. These extensions were only carried out when the transcription initiation peak, and region in between the gene body, had similar coverages of transcription elongation modification across their extent. This class dubbed the "extension class" was further sub-categorized based on the length of extension. Minor extensions being defined as an annotation being increased by less than 500 bp, or the length of a single exon, major extensions defined as increasing the length of annotation between 500 and 2000 bp, and hyper large extension with protein-coding genes needing to be extended upwards of 2000 bp. Novel annotations were classified as those regions with a corresponding transcription elongation peak, as well as a corresponding transcription initiation peak that did not overlap within any known protein-coding, or noncoding gene. Finally, the merged class of annotations were those in which extension caused overlap with another annotation. At these loci the coordinates were shifted to encompass both annotations.

We avoided utilizing this method to split annotations due to the possibility of potential "split" annotations representing separate isoforms of the same transcriptional unit. Due to the hereteogeneous nature of cell types within plant tissues, the aggregate ChIP-seq signal would not provide clear evidence of variable isoforms versus two separate transcriptional units.

Tandem duplication analysis

To test for tandem duplicates in the merger class, we generated a blast protein database containing the original protein-coding sequences of all genes found in this class. Tandem duplicates were defined as those which had a percent identity greater than 50%, and could align to at least 50% of the query protein sequence length. Dotplots were also generated for all pairs, and manually inspected for obvious signs of duplication. In addition, 68 gene pairs out of the 363 were removed from this analysis, as we identified a set of annotations with multiple genes annotated which were completely overlapping, and annotated to the same transcriptional start sites. These loci likely represent different isoforms of the same gene which have been misannotated, and were thus discarded from our tandem duplicate analysis.

Assembly and validation of updated annotations in maize

To validate updated loci, reads overlapping the hypothesized annotation regions were pulled from 23 strand specific tissue types of the maize tissue atlas (Walley *et al.* 2016). StringTie was used to assemble transcripts in each region with parameters "–rf -f 0.01 -a 2 -m 50 -c 3.0 -f 0.0." Updated transcripts were then compared to old annotations, and categorized as correct if the updated transcript was larger than the original annotations. For further validation, Iso-seq reads were gathered from three different array express projects E-MTAB-7837, E-MTAB-7394, E-MTAB-3826, E-MTAB-5957, E-MTAB-5915, and E-MTAB-5956, aligned using STARlong "–outFilterMultimapScoreRange 1 –outFilterMismatchNmax 2000 –winAnchorMultimapNmax 200 –scoreGapNoncan -20 –scoreGapGCAG -4 –scoreGapATAC -8 –scoreDelBase -1 –scoreDelOpen -1 –scoreInsOpen -1 –scoreInsBase -1 –seedSearchLmax 30 –seedSearchStartLmax 50 –seedPerReadNmax 100000 –seedPerWindowNmax 1000 –alignTranscriptsPerReadNmax 100000 –alignTranscriptsPerWindowNmax 10000" (Dobin *et al.* 2013; Wang *et al.* 2016, 2018, 2020). Predicted annotation regions were compared to Iso-seq alignments, and regions that had a corresponding Iso-seq alignment which was greater than the original annotation were considered as passing. Transcripts were further processed through Transdecoder to identify open reading frames, with the following parameters being used "TransDecoder.LongOrfs -m 50" and "TransDecoder.Predict –retain_long_orfs_length 100" (https://github.com/TransDecoder, Last accessed 8.2.2021) .

Reannotation of other species using chromatin data

Histone modification data were downloaded from a list of seven species from previous work; gene expression omnibus number GSE128434 (Lu *et al.* 2019). Reads were downloaded from GEO, and treated identically as outlined in the ChIP-seq section of the methods. Identical read alignment, and peak calling were performed, adjusting for relative genome size in the epic2 and MACs2 to alter peak calling stringency (Zhang *et al.* 2008 p. 2; Stovner and Sætrom 2019). No replicates existed for other species.

Results

To determine if histone modification data could be leveraged to improve genome annotations, we used previously published ChIP-seq data of histone modifications from leaf, root, and inflorescence tissue of *Zea mays* (Ricci *et al.* 2019), which is known to be challenging to annotate (Wang *et al.* 2016). These data were used to define the chromatin landscape around expressed genes. As previously reported, these data showed the expected enrichment of histone modifications around expressed genes, with H3K36me3 and H3K4me1 occurring across the gene body of

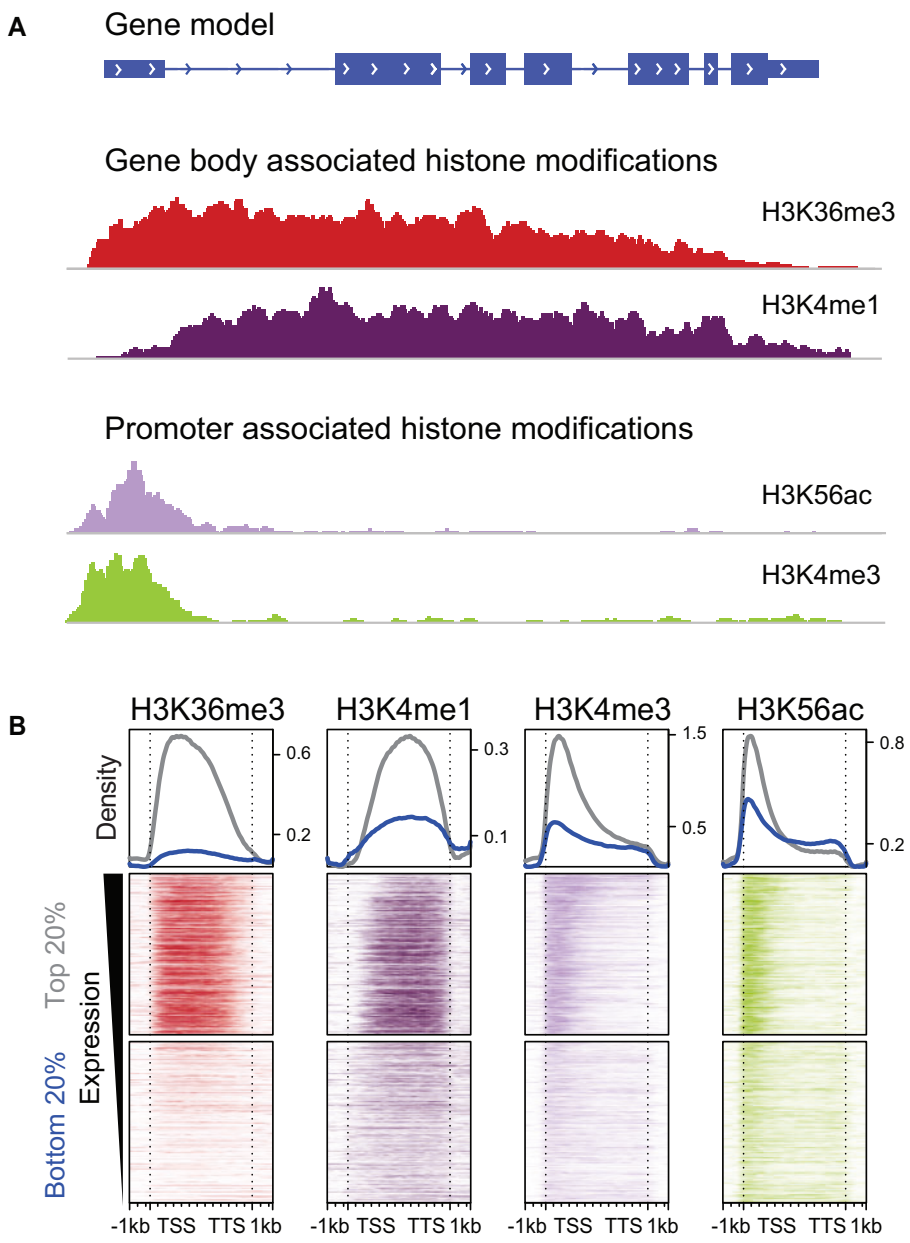

Figure 1 The distribution of histone modifications across expressed genes: (A) An example of an annotated gene body, with corresponding histone modifications indicative of transcription. Histone modifications associated with transcription initiation (H3K56ac and H3K4me3) are known to correspond to the promoters, whereas histone modifications associated with transcription elongation (H3K36me3 and H3K4me1) occur across the length of the gene body. (B) Metaplots of the top and bottom 20% of expressed genes in the genome of *Z. mays* ordered by expression.

actively transcribed genes as indicated by RNA-seq data, and histone modifications H3K4me3 and H3K56ac at promoters (He *et al.* 2003; Kizer *et al.* 2005; Chunyan *et al.* 2010) (Figures 1 and 2A). Hereafter, we refer to these histone modifications collectively as either "transcription initiation" (H3K56ac and H3K4me3) or "transcription elongation" (H3K4me1 and H3K36me3). These histone modifications together are representative of the chromatin environment around transcribed genes, considering many other modifications correlate with those chosen (Li *et al.* 2007; Berr *et al.* 2011). We evaluated the co-occurrence of regions enriched for either histone modification comprising transcription initiation, and found that 64% of these regions co-occurred, as compared to 74% of gene body (H3K36me3 and H3K4me1) histone modifications (Figure 2B). The percentage co-occurrence between similar histone modifications is consistent across additional sampled tissues: root and inflorescence (Supplementary Figures S1 and S2).

A greater number of transcription initiation enriched regions than transcription elongation enriched regions were found in all tissues sampled (7719 excess enriched domains in leaf, 9327 in inflorescence, and 12,164 in root). This larger number of transcription initiation enriched domains as compared to transcription elongation modifications can be explained by multiple reasons. In total, 19,724 transcription initiation regions in leaf, 23,387 in root, and 20,941 in inflorescence overlapped genes. This discrepancy compared with the transcription elongation modifications can be in part explained by the fact that 917 genes in leaf, 1580 in root, and 1331 in inflorescence overlapped greater than one transcription initiation enriched regions, totaling an additional 1981 transcription initiation domains in leaf, 3235 in

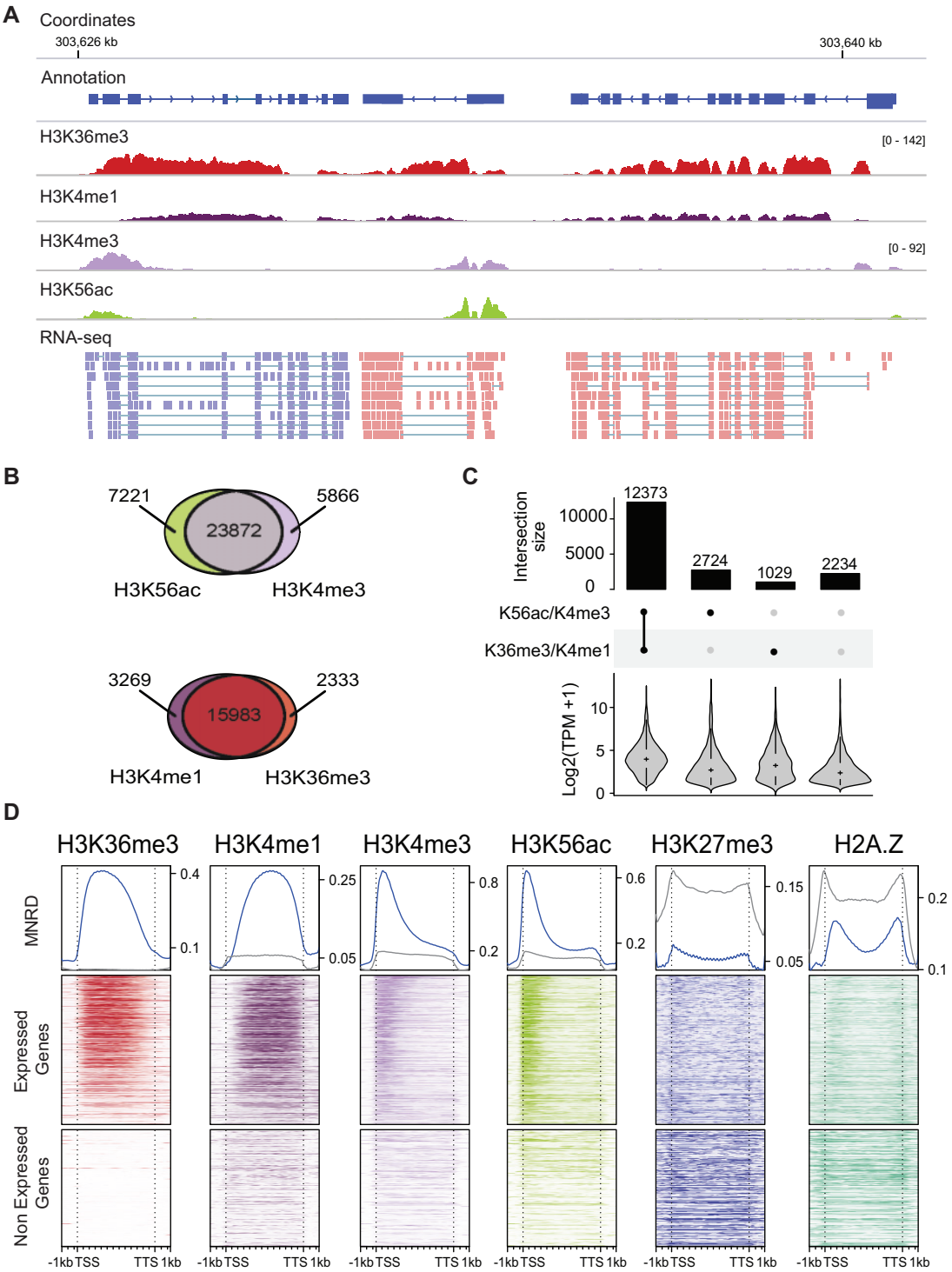

Figure 2 The histone modification landscape of expressed and unexpressed genes in *Z. mays*. (A) Representative screenshot of expressed genes in the genome, and corresponding histone modifications. The histone modifications H3K36me3 and H3K4me1 are found across gene bodies, whereas H4K4me3 and H3K56ac are found in promoters of genes. (B) Venn diagrams of enriched domains for co-occurring histone modifications that reflect either transcription initiation or transcription elongation in the maize leaf. (C) Upset plot with number of genes with greater than 1 TPM, and their intersection with histone modifications associated with either transcription elongation and transcription initiation. (D) Metaplot profiles of expressed and unexpressed genes.

root, and 2736 in inflorescence. In addition, of genes that overlapped transcription initiation modifications, 2584 in leaf, 3441 in root, and 2107 in inflorescence overlapped H3K27me3, a known repressive histone modification. A total of 4822 in leaf, 6018 in root, and 5939 transcription initiation regions did not overlap with any annotated gene. Interestingly, of the subset of these transcription initiation modifications; 550 in leaf, 805, in inflorescence, and 752 in root overlap H3K27me3 domains, possibly indicating silenced unannotated genes in the genome. In addition, 1669 transcription initiation enriched loci in leaf, 2795 in inflorescence and 1782 in root, overlapped a region also enriched for at least one transcriptional elongation histone modification,

possibly representing a set of unannotated genes. Finally, a subclass of 2340 inflorescence, 2412 leaf, and 3486 root transcription initiation enriched regions show no overlap with transcription elongation modifications (Supplementary Figure S3). These regions are generally small with a mean size of 678 bp, and are on average 42,088 bp away from the nearest gene (Supplementary Figure S3). The exact function of these regions remains unknown, but comparative epigenomic approaches will be useful to further understand them in the future.

The concordance of histone modification data around expressed protein-coding genes was used to evaluate their potential to identify actively transcribed regions of the genome. Genes that had a transcript per million (TPM) value greater than 1 were labeled as "active" whereas those which had a TPM value less than 1 TPM were labeled as "inactive." To ensure that the analysis did not suffer from *in silico* biases created by mappability issues in the maize genome, only genes that were greater than 70% mappable were used for analysis (see Methods). Overall, 67% of active genes had both histone modifications indicative of transcription initiation and elongation (Figure 2C). Genes that had both histone classes were likely to be more highly expressed as compared to the other three groups (harboring only one class of histone modification, as compared to two, or no domain enrichment), a trend observed across all three tissue types examined (Kolmogorov-Smirnov tests: $P < 2.2e\text{-}16$) (Figure 2C; Supplementary Figures S1 and S2).

To further demonstrate the relationship between histone modifications and transcribed regions of the genome, we evaluated the distribution of the histone modifications throughout gene bodies of active and inactive genes by generating metaplots (Figure 2D). Transcribed genes generally show enrichment for histone modifications of both transcription initiation, as well as transcription elongation. Active genes also display the expected meta profiles of the sampled histone modifications, with H3K36me3 showing increased enrichment at the 5' region of gene bodies, and H3K4me1 showing increased enrichment at the 3' end. In contrast, inactive genes show no enrichment for transcriptionally related histone modifications, but do show enrichment for histone modifications (H3K27me3) and variants (H2A.Z) associated with facultative heterochromatin (Luo and Lam 2010). These modifications are well documented to be present in genes silenced by polycomb repressive groups of proteins, generally demarcating developmental or environmental specific genes (Coleman-Derr and Zilberman 2012). The slight enrichment in H3K4me3 around these silenced genes likely represents a set of genes bivalently modified, likely poised for rapid upregulation (Zeng *et al.* 2019). These results are similar for both inflorescence and root tissues as well, and are consistent with expectations based on previous findings about the distribution of histone modifications around active and inactive genes (Supplementary Figures S1 and S2).

In the analysis of the histone modifications around expressed genes, we identified two distinct subclasses of genes which violated the expected distributions of histone modifications. One such subset of genes only co-occurred with transcription elongation histone modifications, whereas the other exclusively co-occurred with transcriptional initiation histone modifications (Figure 2C). After manually inspecting a set of genes from each of these classes, we realized that a substantial proportion could be explained by misannotations, with the histone modification data clearly denoting the true extent of the gene model. For instance, in the transcriptional elongation only class, oftentimes the correct transcription initiation start site was clearly evident directly upstream. This led us to speculate that histone modification data can be leveraged to improve gene annotations and to identify novel genes not previously annotated in the genome.

Identification of previously ambiguous annotation classes

After manually inspecting regions of the genome where the histone modification data was discordant with the annotation, we identified three distinct classes of putative misannotations. One class labeled the "Gene merger" class featured histone modification data supporting a single transcriptional unit, but instead, multiple gene annotations existed at these loci in the reference (Figure 3A). Furthermore, alignment of RNA-seq data clearly shows reads bridging the gap between many of these putative misannotations, further supporting that these are a single transcribed unit. A second class of annotation issues found was regions of the genome that had evidence of transcription, and yet had no annotation present in the reference annotation. This class, labeled the "Novel class," likely identifies novel protein-coding genes or lncRNAs (Figure 3B). Finally, we identified an annotation class based off of missing downstream or upstream regions of the transcribed unit that we labeled the "Extension class" (Figure 3C). This annotation class is defined by signals of transcription initiation appearing upstream of the annotation, or transcription elongation histone modifications extending past their current length of the full transcript.

We further subdivided the "Extension class" based off the distance added to the original annotation, with minor extensions being annotations which were only extended by less than 500 bp or the length of a single exon, major extensions comprising regions falling between 500 and 2000 bp, and hyper large extension being those greater than 2000 bp.

Using these defined classes, we implemented a method to identify these regions genome wide (see Methods section), across three different maize tissues (inflorescence, leaf, and root). In total, we identified 4004 potential novel annotations, with 66% (2645 loci) being identified in only a single tissue. We found 363 potential gene merger events, with 166 (45%) of these mergers being found in all tissues sampled. Of the potential mergers, 357 (98.3%) of the predicted mergers consisted of gene pairs, with the remaining six (1.65%) representing loci where three or more genes were hypothesized to be a single transcriptional unit, in total encompassing 732 gene features. Furthermore, 108 (29.8%) of the potential merger events have identical gene ontology (GO) terms, possibly indicating a single locus which was divided into two during the annotation process. To rule out potential assembly errors being the main cause of this merger class, we intersected our merger class with a list of B73 contigs, and found all but one (99.72%) were found on a single contig, ruling out large scale genomic assembly errors as a potential cause of these merged genes. In addition, to ensure that this approach was not merging tandem gene duplicates, we used BLASTP to compare the protein-coding sequences of merged pairs, and looked for sequence identity (States and Gish 1994). We found that only four (1.4%) of these merged pairs had any significant sequence identity between them and of these four pairs, only two had identical GO terms. Finally, of the three extension classes, 4252 minor extensions, 4064 major extensions, and 543 hyper large extensions were found. For both major, and minor extensions, root comprises the highest proportion of uniquely identified extensions, comprising 17% of the major extension class and 31% of

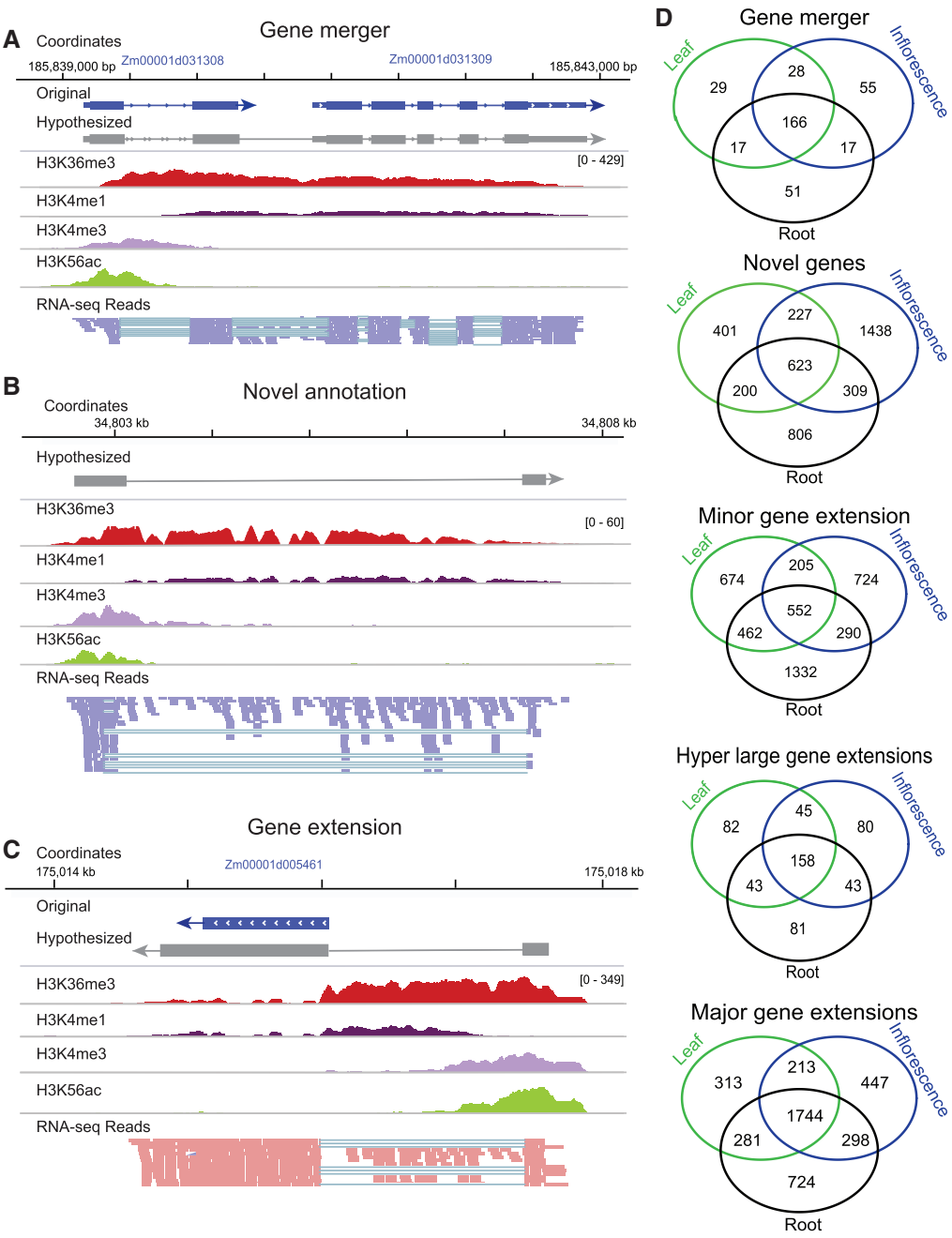

Figure 3 Representative examples and counts of histone modification discordant annotations. (A–C) Representative examples of annotation types found. Current annotations are represented in blue, and the hypothesized annotations are in gray. Histone modification data on the bottom coincides with the length of the gene body (H3K36me3 or H3K4me1) or with the transcription start site (H3K4me3 and H3K56ac). (D) The number of each annotation class found in one of the three tissues sampled (leaf, root, and inflorescence).

the minor extension class. Transcripts found in each annotation class were additional scanned for functional domains, we found that within the hyper large gene class 433 (80%) had a functional domain, as compared to 441 (60%) genes in the merger class, 2627 (61%) in the minor extension class, and 2944 (72%) in the major extension class. In total, using histone modification data we were able to identify 13,159 loci requiring further investigation. Either encompassing misannotations or potential novel loci which have gone unannotated until now. With these regions identified, we were then interested to see if we could validate these hypothesized annotations.

Validation of hypothesized annotations

After identifying putative misannotations, we sought to validate these hypothesized annotations by reassembling transcripts at the specified locus using more inclusive computational parameters. In parallel to assembling transcripts from RNA-seq data, we also utilized full-length transcript isoform sequencing using PacBio Iso-seq reads from multiple studies (see Methods) to evaluate hypothesized annotations. Overall, 67% (335) of the hyper large class of genes were validated by both long-read sequencing, as well as re-assembled transcripts from short reads, and 21.5% (115) were supported by one of these data types (Figure

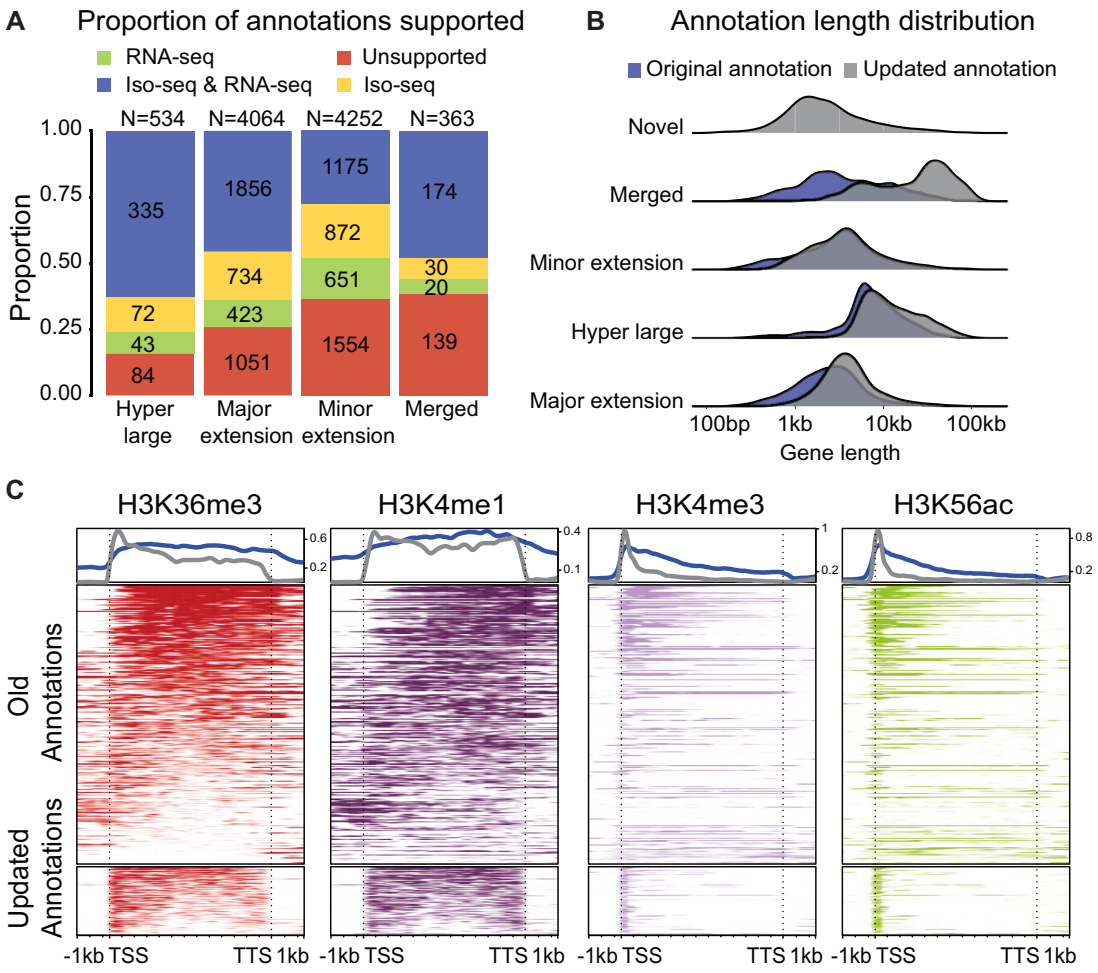

Figure 4 Validation of hypothesized annotations. (A) The proportion of annotation classes validated by either long-read sequencing data, reassembled RNA-seq data, both, or neither (B) Distribution of each annotation class before and after updated annotations were generated. Only annotations that were validated by at least one supplementary data source were used for further analysis. (C) Metaplots of histone modifications surrounding the merger class of updated annotation regions. The blue lines in the profile represent the annotations before reassembly, and the gray lines represent the annotations after.

4A). For the major extension class, 45.7% (1856) of regions were supported by both RNA-seq and Iso-seq reads, with 28.5% (1157) being validated by a single data type, and 25.9% (1051) of major extension annotations being unsupported. In the gene merger class, 47.9% (174) of the hypothesized mergers were validated with RNA-seq and Iso-seq, 13.77% (50) supported by a single data type, and 38.9% (139) had no additional support. In total, 68% (2698) of the minor extension class were validated by at least one alternative data source. For the novel class of annotations, we re-assembled regions using RNA-seq from the corresponding tissue in which the novel region was identified. In total, 72% (3253) of the novel loci were supported by an assembled transcript. Of these transcripts, we found that 74% (2421) were able to generate an open reading, indicating that the remaining 28% likely constitute lncRNAs. Overall, 6385 out of 9213 of the potential misannotations were found to be corroborated from orthogonal datasets, demonstrating the capacity of histone modification data to allow for identification of potentially misannotated regions, and hypothesis-driven annotation correction.

We next evaluated how the distribution of gene length shifted after reannotation for each class. Only loci which had at least one type of supplementary support (Iso-seq reads or RNA-seq) were used for this analysis (Figure 4B). Overall, the distribution of the merged class was the most radically changed, as the median gene size shifted from 3089 bp in length to 29,704 bp (Supplementary Figure S4). In contrast, the median gene size for the novel class is 1962 bp, smaller than the median known gene size for maize which is 2568 bp (Portwood *et al.* 2019). The major extension gene class shifted from a median size of 2363 to3818 bp, the minor extension class shifted from 3165 to 3631 bp, and the hyper large gene class shifted from 6838 to 10,909 bp. To determine if the re-annotated regions more accurately recapitulated the expected distribution of histone modifications around a transcribed gene, we regenerated metaplots. The updated annotation sets more accurately reflect the known landscape of histone modifications around transcribed units (Figure 4C). This trend appears similarly in all found annotation classes (Supplementary Figures S5 and S6). This implementation of histone modification data allowed us to recapture previously unannotated regions in the genome of *Z. mays* while also improving existing annotations. All updated annotation coordinates are found in Supplementary Table S1. In addition, we compared the class of merged annotations against a known list of 78 split annotations pairs available on Gramene (Tello-Ruiz *et al.* 2021). In total, 31 (40%) of the Gramene split annotations were concordant with the annotation mergers identified by our methods. The

remaining 47 split gene pairs were either in regions where there was missing data (20/78), mappability issues (20/78), or were missed due to complex loci with multiple gene features in diverging directions (7/78). Comparisons between the merged dataset and the Gramene split gene dataset are in Supplementary Table S2. Although the Gramene split list is a set of well-documented split errors in the maize genome, more recent studies using comparative annotation-based approaches have also been implemented, posing an excellent opportunity to compare and contrast the identified list of gene merger pairs further.

We were interested in comparing the merged annotation group against a recent study that aimed to improve annotation of the maize genome by comparing annotations of numerous *Z. mays* cultivars (B73, PH207, and W22) against one another (Monnahan *et al.* 2020). By utilizing a blastp based approach for identification of potential gene merger pairs, followed by an analysis focused on variation in expression patterns across tissues, they identified split gene pairs that should be merged across the genome (Monnahan *et al.* 2020). In total, 109 (48%) of the merged annotation class identified in this study intersected gene merger pairs identified in the Monnahan *et al.* study. Out of these 109 cross captured merger pairs, 34 were represented in the high confidence gene merger class identified in Monnahan et al. In addition, 60 out of the 109 (55%) of the mergers found at the intersection of our studies fall into instances where they were identified in Monnahan *et al.*, but unable to be confidently classify based off of differential RNA-seq analysis. The histone modification data in our study provides clear evidence independent of RNA-seq that these 60 loci should be merged (Supplementary Figure S7). Finally, there was a small class of 15 loci (14%) that were discordant between the two methods. With the histone modification data supporting gene merger, whereas the analysis by Monnahen *et al.* identifies that these loci should remain as split pairs. All intersecting annotations found between our studies are in Supplementary Table S2. Overall, the concordance between these gene merger sets demonstrates the inherent challenge associated with genome annotation while also demonstrating the advantage ChIP-seq provides as an orthogonal assay to RNA-seq based methods for gene annotation.

Knowing that Monnahan *et al.* identified 96 high-confident gene merger pairs, we were interested in further investigating the remaining 62 pairs to ascertain why these potential misannotations were not identified by our method. Of the remaining 62 high confidence candidates identified by Monnahan *et al.*, the histone modification data indicates that 30 of them should not be merged (Supplementary Figure S8). The data provided by ChIP-seq provides strong evidence of distinct genes possessing their own transcription start sites and evidence of unique transcriptional elongation activities at each gene (Supplementary Figure S8). Upon individual inspection of the remaining uncaptured 32 high confidence merger pairs identified by Monnahan *et al.* these candidates existed in either low mappability regions of the genome (10/32) or did not intersect a combination of histone modification enriched domains within the tissues that we sampled (22/32). The lack of being able to capture these loci is a limitation of our method; demonstrating the essentiality of utilizing many methods to improve genome annotations.

Reannotation of multiple plant genomes

After successfully applying this method in *Z. mays*, we were interested in extending this method to other plant genomes with available high-quality histone modification data. In total, we included an additional five species, *Asparagus officinalis*, *Setaria viridis*, *Sorghum bicolor*, *Glycine max*, and *Phaseolous vulgaris* (Lu *et al.* 2019). In total, we identified 4640 novel annotations present in the *A. officinalis*, 3404 minor extensions, 386 potential gene mergers, 3090 major extensions, and 2316 hyper large extensions (Figure 5A). The abundance of potential novel transcripts identified in *A. officinalis* was unexpectedly high, given that only a single tissue type, leaf, was used for this analysis. This stands in contrast to *Z. mays*, where across the three tissue types sampled, we found 4004 potential novel regions. In *G. max*, we found a

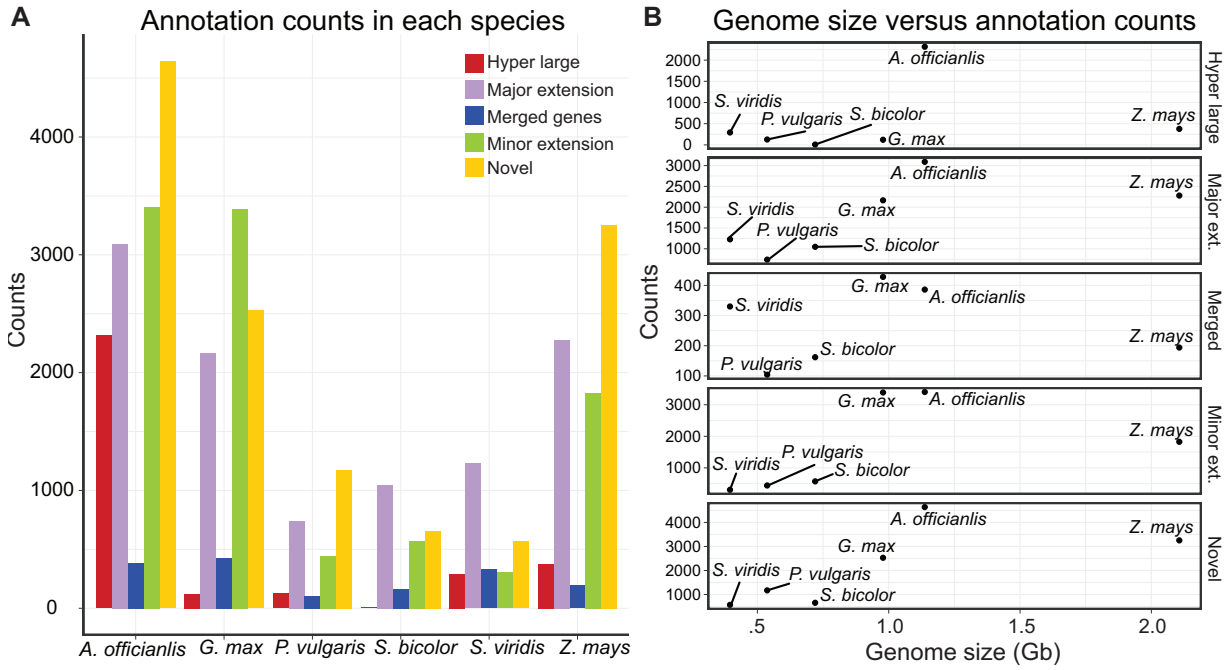

Figure 5 Reannotation of diverse plant genomes using epigenomic data. (A) Counts of each annotation type identified in each plant genome. (B) Scatterplot of genome size of each species annotated versus the number of annotation counts in each class.

further 121 hyper large extensions, 2165 major extensions, 3388 minor extensions, 428 merged genes, and 2529 novel annotations. The annotations and relative counts of each annotation class found in each species are found in Supplementary Table S2. This analysis clearly demonstrates that histone modification data can be utilized on diverse plant genomes to quickly assay the quality of genomic annotations in a given tissue type.

Upon generating a list of hypothesized annotations, we noticed a slight trend in regards to genome size and putative annotation errors. We noticed that the smaller genomes that we sampled, namely *P. vulgaris, S. bicolor, and S. viridis* appeared to have smaller number of potential genome annotation errors as compared to larger genomes. By correlating genome size with the counts of annotation error of each type, we found that larger genomes have more errors in the extension and novel classes of genes (Figure 5B). Although this trend appears to be true for *G. max, and A. officinalis*, it breaks down for the largest genome sampled here, *Z. mays*. However, the fact that *Z. mays* does not continue this trend may reflect the attention that this plant garners; with less annotation errors reflecting the abundance of resources, and groups working it. This large proportion of hyper large and major extension genes also appears to reflect a certain level of bias when annotating plant genomes, as we capture more issues in regards to large gene classes in plant genomes which are larger, and likely have a history of transposon expansion around gene features, causing increased intron size (Figure 5B).

Discussion

In the post-genome assembly era, annotation represents the next great hurdle in accurate genomic resource creation. Here, we demonstrate that histone modification data offers a valuable untapped resource to precisely improve plant genome annotations. By easily assessing the transcribed space of the genome and identifying domains enriched with histone modifications that correlate with specific transcriptional events, valuable hypotheses about annotation features can be generated. These hypotheses, such as identifying potential transcript length and location of transcription start sites, can be used in a manner complementary to RNA-based methods to provide a way to quickly fix gene models, and generate more accurate genome annotations.

This study demonstrates the power and advantages of using histone modification data to generate hypothesis about the transcribed genic space, offering valuable orthogonal assay. By utilizing histone modifications on a genome-wide scale, we identified consistent trends where annotations were discordant with the expected distribution of histone modification data and identified five distinct classes of annotation errors. We validated a set of these annotations using RNA-based methods. In total, we were able to identify, and validate 7930 annotation errors. Of these updated transcripts, 3253 represent novel transcripts, demonstrating the capacity of histone modification data to capture previously unannotated genes. Upon correction and reannotation, these updated annotations more accurately reflected what is known about the histone modification landscape of transcribed genes and captured previously unannotated gene space.

In addition, this study shows that the usefulness of epigenomic data is not unique to *Z. mays*. To demonstrate this we assayed five additional plant genomes for possible annotation errors using this method, and found varying abundances of either novel or misannotations. We correlated the counts of annotation errors with genome size, and found a slight correlation between the two, although additional studies of a great number of species will be required to know if this is significant. The abundance of potential annotation errors found across these five species demonstrates the importance of having orthogonal support for gene annotations and illustrates the challenges in making accurate *a priori* assumptions about gene features in plants.

Annotation errors are a natural part of generating genomic resources. The complexity of genic space, paired with the tissue and cell type specificity of many genes, and the assumptions required in each *in silico* step of annotation converge to create an exceptionally challenging problem. These myriad challenges make annotation errors an inevitability, and downstream curation a necessity. Currently, sophisticated community-driven approaches exist to identify and fix annotation errors, but these large-scale efforts are limited to only well-studied species. This bias in community size greatly inhibits the potential value, as the species with assembled genome increase in diversity.

The methodology presented here offers a protocol to appraise current annotations, and potentially fill in this downstream gap. However, while valuable, it is important to note that this method is not a panacea. ChIP-seq remains a challenging experiment and is not used as frequently as compared to RNA-seq. The lack of publicly available data, as well as the limited number of the tissue types sampled diminishes utility of this method. However, the increased accuracy added to genome annotations due to this method certainly introduce the potential of ChIP-seq becoming a standard protocol when considering genome annotation methods. Having a sequenced genome is only the first step to creating a valuable biological resource, and the challenges facing the production of accurate genome annotations remain. Epigenomic data offers one powerful orthogonal resource which, when utilized correctly, can strengthen current efforts and mitigate some, but not all, issues of genomic annotation moving forward.

Data availability

All novel data generated for this analysis can be found under the GEO accession number GSE160944. The code used to run the above analysis can be found on GitHub in the following repository, https://github.com/Jome0169/MendietaPablo_Annotation_Paper_scripts. Of special interest is the script Update_annotation.py, which implements the re-annotation pipeline discussed in the method section. Updated annotations and gene models for *Z. mays* can be viewed at the Plant Epigenome JBrowse Genome Browser. Supplementary material is available at figshare: https://doi.org/10.25387/g3.14885733.

Acknowledgments

The author thanks all the Schmitz lab members for their consistent feedback, and special thanks to Katie Duval for her willingness to look over multiple rounds of figure generation and editing. J.P.M. and R.J.S. led the conceptualization of this project. J.P.M. led the analysis and writing of this manuscript. A.M. contributed assistance in editing figures, as well as provided valuable feedback throughout the study. X.Z. and W.A.R. both contributed ChIP-seq experiments.

Funding

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National

Institute of Health under award number T32GM007103. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. A.M. was supported by the NSF Postdoctoral Fellowship in Biology (DBI-1905869). This study was funded by support from the National Science Foundation (IOS-1856627) and the UGA Office of Research awarded to R.J.S.

Conflicts of interest

R.J.S. is a co-founder of REquest Genomics, LLC, a company that provides epigenomic services. J.P.M, X.Z, A.P.M, and W.A.R declare no conflict of interest.

Literature cited

Allfrey VG, Faulkner R, Mirsky AE. 1964. Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. Proc Natl Acad Sci USA. 51:786–794.

Alvarez-Venegas R, Pien S, Sadder M, Witmer X, Grossniklaus U, *et al.* 2003. ATX-1, an Arabidopsis homolog of trithorax, activates flower homeotic genes. Curr Biol. 13:627–637.

Bannister AJ, Kouzarides T. 2011. Regulation of chromatin by histone modifications. Cell Res. 21:381–395.

Bernatavichute YV, Zhang X, Cokus S, Pellegrini M, Jacobsen SE. 2008. Genome-wide association of histone H3 lysine nine methylation with CHG DNA methylation in *Arabidopsis thaliana*. PLoS One. 3:e3156.

Bernstein BE, Kamal M, Lindblad-Toh K, Bekiranov S, Bailey DK, *et al.* 2005. Genomic maps and comparative analysis of histone modifications in human and mouse. Cell. 120:169–181.

Berr A, Shafiq S, Shen WH. 2011. Histone modifications in transcriptional activation during plant development. Biochim Biophys Acta. 1809:567–576.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 30:2114–2120.

Bu Z, Yu Y, Li Z, Liu Y, Jiang W, *et al.* 2014. Regulation of Arabidopsis flowering by the histone mark readers MRG1/2 via interaction with constants to modulate FT expression. PLoS Genet. 10: e1004617.

Cartagena JA, Matsunaga S, Seki M, Kurihara D, Yokoyama M, *et al.* 2008. The Arabidopsis SDG4 contributes to the regulation of pollen tube growth by methylation of histone H3 lysines 4 and 36 in mature pollen. Dev Biol. 315:355–368.

Cazzonelli CI, Cuttriss AJ, Cossetto SB, Pye W, Crisp P, *et al.* 2009. Regulation of carotenoid composition and shoot branching in Arabidopsis by a chromatin modifying histone methyltransferase, SDG8. Plant Cell. 21:39–53.

Chunyan L, Falong L, Xia C, Xiaofeng C. 2010. Histone Methylation in higher plants. Annu Rev Plant Biol. 61:395–420.

Coleman-Derr D, Zilberman D. 2012. Deposition of histone variant H2A.Z within gene bodies regulates responsive genes. PLoS Genet. 8:e1002988.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, *et al.* 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 29:15–21.

Dozmorov MG. 2017. Epigenomic annotation-based interpretation of genomic data: from enrichment analysis to machine learning. Bioinformatics. 33:3323–3330.

Earley KW, Shook MS, Brower-Toland B, Hicks L, Pikaard CS. 2007. *In vitro* specificities of Arabidopsis co-activator histone acetyltransferases: implications for histone hyperacetylation in gene activation: Specificities of Arabidopsis HATs. Plant J. 52:615–626.

Ernst J, Kellis M. 2017. Chromatin state discovery and genome annotation with ChromHMM. Nat Protoc. 12:2478–2492.

Fromm M, Avramova Z. 2014. ATX1/AtCOMPASS and the H3K4me3 marks: how do they activate Arabidopsis genes? Curr Opin Plant Biol. 21:75–82.

Gu Z, Eils R, Schlesner M, Ishaque N. 2018. EnrichedHeatmap: an R/Bioconductor package for comprehensive visualization of genomic signal associations. BMC Genomics. 19:234.

Guttman M, Amit I, Garber M, French C, Lin MF, *et al.* 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 458:223–227.

Hannah A. 1951. Localization and function of heterochromatin in *Drosophila Melanogaster*. In: Advances in Genetics. 4:87-125. https://www.sciencedirect.com/science/article/pii/S0065266008602321.

He Y, Michaels SD, Amasino RM. 2003. Regulation of flowering time by histone acetylation in Arabidopsis. Science. 302:1751–1754.

Jarroux J, Morillon A, Pinskaya M. 2017. History, discovery, and classification of lncRNAs, In: MRSRao, editor Long Non Coding RNA Biology. Advances in Experimental Medicine and Biology. Singapore: Springer Singapore, p. 1–46.

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.

Jin J, Shi J, Liu B, Liu Y, Huang Y, *et al.* 2015. MORF-RELATED GENE702, a reader protein of Trimethylated Histone H3 Lysine 4 and Histone H3 Lysine 36, is involved in Brassinosteroid-regulated growth and flowering time control in rice. Plant Physiol. 168:1275–1285.

Kizer KO, Phatnani HP, Shibata Y, Hall H, Greenleaf AL, *et al.* 2005. A novel domain in Set2 mediates RNA polymerase II interaction and couples Histone H3 K36 Methylation with Transcript Elongation. Mol Cell Biol. 25:3305–3316.

Langmead B, Salzberg SL. 2012. Fast Gapped-Read Alignment with Bowtie 2. Nat Methods. 9:357–359.

Li B, Carey M, Workman JL. 2007. The role of chromatin during transcription. Cell. 128:707–719.

Li Q, Gent JI, Zynda G, Song J, Makarevitch I, *et al.* 2015. RNA-directed DNA methylation enforces boundaries between heterochromatin and euchromatin in the maize genome. Proc Natl Acad Sci USA. 112:14728–14733.

Li X, Wang X, He K, Ma Y, Su N, *et al.* 2008. High-resolution mapping of epigenetic modifications of the rice genome uncovers interplay between DNA methylation, histone methylation, and gene expression. Plant Cell. 20:259–276.

Liu J, Seetharam AS, Chougule K, Ou S, Swentowsky KW, *et al.* 2020. Gapless assembly of maize chromosomes using long-read technologies. Genome Biol. 21:121.

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.

Lu C, Tej SS, Luo S, Haudenschild CD, Meyers BC, *et al.* 2005. Elucidation of the small RNA component of the transcriptome. Science. 309:1567–1569.

Luger K. 1997. Crystal structure of the nucleosome core particle at 2.8 A° resolution. Nature. 389:10.

Luo C, Lam E. 2010. ANCORP: a high-resolution approach that generates distinct chromatin state models from multiple genome-wide datasets. Plant J. 63:339–351.

Mahrez W, Arellano MST, Moreno-Romero J, Nakamura M, Shu H, *et al.* 2016. H3K36ac is an evolutionary conserved plant Histone modification that marks active genes. Plant Physiol. 170:1566–1577.

McClintock B. 1950. The origin and behavior of mutable loci in maize. Proc Natl Acad Sci USA. 36:344–355.

Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, *et al.* 2020. Telomere-to-telomere assembly of a complete human X chromosome. Nature. 585:79–84.

Monnahan PJ, Michno J-M, O'Connor C, Brohammer AB, Springer NM, *et al.* 2020. Using multiple reference genomes to identify and resolve annotation inconsistencies. BMC Genomics. 21:281.

Morris SA, Rao B, Garcia BA, Hake SB, Diaz RL, *et al.* 2007. Identification of histone H3 lysine 36 acetylation as a highly conserved histone modification. J Biol Chem. 282:7632–7640.

Nislow C, Ray E, Pillus L. 1997. SET1, a yeast member of the trithorax family, functions in transcriptional silencing and diverse cellular processes. Mol Biol Cell. 8:2421–2436.

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.

Pockrandt C, Alzamel M, Iliopoulos CS, Reinert K. 2020. GenMap: ultra-fast computation of genome mappability. Bioinformatics. 36:3687–3692.

Portwood JL, Woodhouse MR, Cannon EK, Gardiner JM, Harper LC, *et al.* 2019. MaizeGDB 2018: the maize multi-genome genetics and genomics database. Nucleic Acids Res. 47:D1146–D1154.

Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, *et al.* 2011. A unique chromatin signature uncovers early developmental enhancers in humans. Nature. 470:279–283.

Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. 2014. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. 42:W187–W191.

Rando OJ. 2012. Combinatorial complexity in chromatin structure and function: revisiting the histone code. Curr Opin Genet Dev. 22:148–155.

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.

Roudier F, Ahmed I, Berard C, Sarazin A, Mary-Huard T, *et al.* 2011. Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. EMBO J. 30:1928–1938.

Saleh A, Alvarez-Venegas R, Avramova Z. 2008. Dynamic and stable histone H3 methylation patterns at the Arabidopsis FLC and AP1 loci. Gene. 423:43–47.

Salzberg SL. 2019. Next-generation genome annotation: we still struggle to get it right. Genome Biol. 20:92.

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.

Schübeler D, MacAlpine DM, Scalzo D, Wirbelauer C, Kooperberg C, *et al.* 2004. The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. Genes Dev. 18:1263–1271.

Shi J, Dawe RK. 2006. Partitioning of the maize epigenome by the number of Methyl Groups on Histone H3 Lysines 9 and 27. Genetics. 173:1571–1583.

Song Z-T, Sun L, Lu S-J, Tian Y, Ding Y, *et al.* 2015. Transcription factor interaction with COMPASS-like complex regulates histone H3K4 trimethylation for specific gene expression in plants. Proc Natl Acad Sci USA. 112:2900–2905.

States DJ, Gish W. 1994. Combined use of sequence similarity and codon bias for coding region identification. J Comput Biol. 1:39–50.

Stovner EB, Sætrom P. 2019. epic2 efficiently finds diffuse domains in ChIP-seq data. Bioinformatics. 35:4392–4393.

Strahl BD, Allis CD. 2000. The language of covalent histone modifications. Nature. 403:41–45.

Tello-Ruiz MK, Naithani S, Gupta P, Olson A, Wei S, *et al.* 2021. Gramene 2021: harnessing the power of comparative genomics and pathways for plant research. Nucleic Acids Res. 49: D1452–D1463.

van Dijk K, Ding Y, Malkaram S, Riethoven J-JM, Liu R, *et al.* 2010. Dynamic changes in genome-wide histone H3 lysine 4 methylation patterns in response to dehydration stress in *Arabidopsis thaliana*. BMC Plant Biol. 10:238.

Wagner EJ, Carpenter PB. 2012. Understanding the language of Lys36 methylation at histone H3. Nat Rev Mol Cell Biol. 13:115–126.

Walley JW, Sartor RC, Shen Z, Schmitz RJ, Wu KJ, *et al.* 2016. Integration of omic networks in a developmental atlas of maize. Science. 353:814–818.,

Wang B, Regulski M, Tseng E, Olson A, Goodwin S, *et al.* 2018. A comparative transcriptional landscape of maize and sorghum obtained by single-molecule sequencing. Genome Res. 28: 921–932.

Wang B, Tseng E, Baybayan P, Eng K, Regulski M, *et al.* 2020. Variant phasing and haplotypic expression from long-read sequencing in maize. Commun Biol. 3:1–11.

Wang B, Tseng E, Regulski M, Clark TA, Hon T, *et al.* 2016. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. Nat Commun. 7:11708.

Xu L, Zhao Z, Dong A, Soubigou-Taconnat L, Renou J-P, *et al.* 2008. Di- and Tri- but not Monomethylation on Histone H3 Lysine 36 marks active transcription of genes involved in flowering time regulation and other processes in *Arabidopsis thaliana*. Mol Cell Biol. 28:1348–1360.

Yandell M, Ence D. 2012. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet. 13:329–342.

Zeng Z, Zhang W, Marand AP, Zhu B, Buell CR, *et al.* 2019. Cold stress induces enhanced chromatin accessibility and bivalent histone modifications H3K4me3 and H3K27me3 of active genes in potato. Genome Biol. 20:123.

Zhang X, Bernatavichute YV, Cokus S, Pellegrini M, Jacobsen SE. 2009. Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in *Arabidopsis thaliana*. Genome Biol. 10:R62.

Zhang X, Henderson IR, Lu C, Green PJ, Jacobsen SE. 2007. Role of RNA polymerase IV in plant small RNA metabolism. Proc Natl Acad Sci USA. 104:4536–4541.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, *et al.* 2008. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 9:R137.

Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW-L, *et al.* 2006. Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. Cell. 126:1189–1201.

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.

Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S. 2007. Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. Nat Genet. 39:61–69.

Communicating editor: M. Hufford