TITLE

Natural methylation epialleles correlate with gene expression in maize

AUTHORS

Yibing Zeng, R. Kelly Dawe*, and Jonathan I. Gent*

*Corresponding authors

kdawe@uga.edu and gent@uga.edu

ABSTRACT

DNA methylation (5-methylcytosine) represses transposon activity and contributes to inaccessible chromatin structure of repetitive DNA in plants. It is depleted from cis regulatory elements in and near genes, but in some genes it is present in the gene body including exons. Methylation in exons solely in the CG context is called gene body methylation (gbM). Methylation in exons in both CG and non-CG contexts is called TE-like methylation (teM). To develop a broader understanding of methylation in maize genes, we utilized recent genome assemblies, gene annotations, transcription data, and methylome data to decipher common patterns of gene methylation. To compare between genomes, we analyzed each data source relative to its own genome assembly rather than the easier but less accurate method of using one assembly as reference for all. We found that teM genes are mainly silent across plant tissues, are limited to specific maize stocks, and exhibit evidence of annotation errors. We used these data to flag all teM genes in the 26 NAM founder genome assemblies (on average 3,693 genes, 9% of total). In contrast to teM, gbM genes are broadly expressed across tissues. We found that they exist in a continuum of CG methylation levels without a clear demarcation between

unmethylated genes and gbM genes. Analysis of expression levels across diverse maize stocks revealed a weak but highly significant positive correlation between gbM and gene expression. gbM epialleles were associated with an approximately 3% increase in steady-state expression level relative to unmethylated epialleles. We hypothesize based on these data that gbM can contribute toward broad and robust gene expression.

INTRODUCTION

DNA methylation is one part of a multilayered chromatin-based method of repressing transcription and accessibility of repetitive DNA in plants. The mode of repression depends in part on the two or three nucleotide sequence context of the methylated cytosine, generally categorized as CG, CHG, and CHH, where H = A, T, or C. DNA methylation is not restricted to repetitive DNA, however. In most flowering plants, a large fraction of genes can have methylation in the CG context (mCG) in exons [reviewed in (Bewick and Schmitz 2017; Muyle *et al.* 2022)]. Such genes are referred to as gene body methylated genes, gbM genes for short. Genes that have TE-like methylation, both mCG and non-mCG, in their exons are referred to as teM genes. The third and most abundant group of genes are unmethylated in their exons and referred to as UM genes. All three methylation groups have signature expression patterns: gbM genes tend to be broadly expressed across tissues, UM genes tend to be tissue-specific, and teM genes tend to be poorly expressed [reviewed in (Bewick and Schmitz 2017; Muyle *et al.* 2022)].

The mCG in gbM genes is maintained by methyltransferases of the MET1 family (Stroud *et al.* 2013). However, there is no dedicated mechanism to establish gbM on genes where it is absent: Genes that have lost gbM in *met1* mutants do not reacquire gbM after MET1 is returned (at least over a period of eight generations (Reinders *et al.* 2009). Instead, establishment of mCG

in gbM genes is hypothesized to occur slowly and infrequently through intermediate teM-like states mediated by spurious DNA methylation by MET1 and chromomethyltransferases of the CMT family (Niederhuth *et al.* 2016; Wendte *et al.* 2019). Although the direct output of CMTs is mCHH or mCHG, they can also lead to mCG by downstream recruitment of MET1 in Arabidopsis (Lyons *et al.* 2022). According to this hypothesis, non-mCG is removed but mCG endures, leading to an epigenetically stable gbM state. Supporting this view are data showing that CMTs can target gbM genes in Arabidopsis (Zhang *et al.* 2020, 2021; Papareddy *et al.* 2021). Expression of Arabidopsis CMT in *Eutrema salsugineum* (which normally lacks both CMT and gbM*)* induces mCG in gene bodies that persists after CMT is removed (Wendte *et al.* 2019).

Since cytosine methylation increases the frequency of G:C to A:T transitions (Ossowski *et al.* 2010), mCG in coding DNA might be harmful rather than beneficial, unless it provides other benefits that outweigh its mutagenic tendency. Gene body methylation is absent from the vast majority of fungal genomes, even in genomes with methylation at repetitive elements (Bewick *et al.* 2019). The existence of gbM in diverse animals, however, is evidence that beneficial functions likely exist. In fact, some insect genomes have extensive gene body methylation but little methylation of repetitive elements (Bewick and Schmitz 2017). Vertebrates have gene body methylation as well as a dedicated mechanism establishing it: recruitment of DNMT3 methyltransferases coupled to trimethylation of histone H3 lysine 36 (H3K36me3) (Baubec *et al.* 2015; Bröhm *et al.* 2022). In plants, H3K36me3 seems to be unrelated to gbM (Wollmann *et al.* 2017). The functional significance of gbM in vertebrates is unclear but may prevent internal transcriptional initiation (Neri *et al.* 2017; Teissandier and Bourc'his 2017). In some cases, gbM also impacts splicing (Yearim *et al.* 2015; Shayevitch *et al.* 2018). Function of

gbM could be compared to the repression of transposons and other repetitive elements by DNA methylation: Although they are strongly methylated in some eukaryotes, they are poorly methylated or not methylated at all in others, including nematodes, fruit flies, yeasts, and honeybees [reviewed in (Schmitz *et al.* 2019)]. Another apt comparison might be the centromeric histone variant CENP-A, which is essential in most eukaryotes but absent in some (Drinnenberg *et al.* 2014).

Whether mCG in gbM genes has a biologically significant effect on steady-state mRNA levels in plants is not clear. Comparisons of gene expression in *E. salsugineum* (without gbM) with *A. thaliana* (with gbM), has yielded conflicting evidence both for and against gbM promoting gene expression (Muyle and Gaut 2019; Bewick *et al.* 2019). Similarly, analysis of genes that lost gbM through *met1* mutation has also yielded conflicting evidence both for and against gbM promoting gene expression (Shahzad *et al.*; Bewick *et al.* 2019). Experiments using methylation data from the Arabidopsis 1001 Genome Consortium (Kawakatsu *et al.* 2016) found evidence of selection on gbM epialleles as well as small increases in expression in gbM over UM epialleles (Shahzad *et al.*; Muyle *et al.* 2021). The fact that gbM genes tend to be expressed more broadly through development than UM genes raises the possibility that gbM may function in stabilizing gene expression across development, with subtle activating (or repressing) functions depending on the cell type (Takuno and Gaut 2012, 2013; Niederhuth *et al.* 2016). It is also possible that gbM has a function unrelated to normal gene regulation–for example, inhibiting ectopic initiation of transcription in gene bodies. The evidence for such a function is mixed in plants (Choi *et al.* 2020; Le *et al.* 2020). One might also speculate functions for gbM unrelated to transcription.

DNA glycosylases, which demethylate DNA through the base excision repair pathway, provide strong evidence that DNA methylation can function in gene regulation. For example, DNA glycosylases can function to activate genes upon bacterial infection (Halter *et al.* 2021), in response to abscisic acid hormone signaling (Kim *et al.* 2019), and in pollen tube development (Khouider *et al.* 2021). In endosperm, DNA glycosylases act on maternal alleles of some genes to cause genomic imprinting, which is essential for endosperm development [reviewed in (Anderson and Springer 2018)]. In the microgametophyte, including mature pollen, demethylation by DNA glycosylases has been proposed to take on a larger role in gene regulation than in sporophytic cells (Borg *et al.* 2021). While TE-like methylation of TEs themselves is generally stable across sporophytic development (Crisp *et al.* 2020), TE-like DNA methylation in coding DNA might indicate a function in gene regulation.

Maize (Zea mays) appears to have typical patterns of gene methylation, including gbM in a set of genes that are typically longer and have broad expression across cell types (Takuno and Gaut 2013; Niederhuth *et al.* 2016; Seymour and Gaut 2020; Martin *et al.* 2021). The promoters of most functional maize genes are constitutively demethylated, and methylation in promoters is a strong indicator of gene silencing (Hufford *et al.* 2021). Recently, improved genome assemblies and annotations were produced for the B73 inbred line and 25 other diverse inbred lines known as the NAM founders (Hufford *et al.* 2021). Transcriptome data for ten tissues and 20X coverage methylome data for developing leaves for each genome provide an opportunity to better characterize gene methylation trends in maize. We made use of this resource to identify natural epialleles and explore the relationships between methylation (UM, gbM, and teM) and gene expression on a pan-genome scale.

Table 1: Usage of key terms

UM	unmethylated in coding DNA sequence (CDS)
gbM	gene body methylation, only mCG in CDS
teM	TE-like methylation, mCG and mCHG in CDS
epiallele type	methylation status of genes: UM, gbM, or teM
tissue-specific gene	expressed in at least one tissue and not expressed in at least one other tissue
constitutive gene	expressed in all ten tissues examined
silent gene	not expressed in all ten tissues
NAM founders	set of 26 diverse maize inbred stocks including B73
core gene	annotated in syntenic position in all NAM founders
pangene	unique identifier for all syntenic homologs of a gene across NAM founders
1-to-1 pangene	pangene with intact tandem duplicates in NAM founders
1-to-N pangene	pangene without intact tandem duplicates in NAM founders
stable pangene	represented by only one epiallele in NAM founders
unstable pangene	represented by more than one epiallele in NAM founders

RESULTS

UM and gbM genes are part of a continuum, teM genes form a distinct group

We first surveyed the landscape of gene methylation in maize using the B73 genome and a DNA methylome from developing seedling leaves as a reference (Hufford *et al.* 2021). Introns often have TE-like methylation that is distinct from flanking exons simply because they contain TE insertions (Seymour and Gaut 2020). The very 5' and 3' end of UTRs are typically unmethylated when they are correctly annotated, but UTR annotations are often imprecise and sometimes overlap with nearby TEs. Thus we excluded both UTRs and introns in measuring genic methylation. For each gene with sufficient read coverage, we assigned a single mCG value and mCHG value as the average methylation in its coding DNA sequence (CDS). To produce a visual summary of gene methylation trends, we represented each gene by these two values. A clear bimodal trend was evident, where the larger group of genes had low mCHG methylation and a continuous range of mCG but heavily skewed toward near zero mCG (Fig. 1A). A second, smaller group of genes had both high mCG and mCHG (Fig. 1A). We divided the first group into UM genes and gbM genes, where UM had less than or equal to 0.05 mCG and gbM had at least

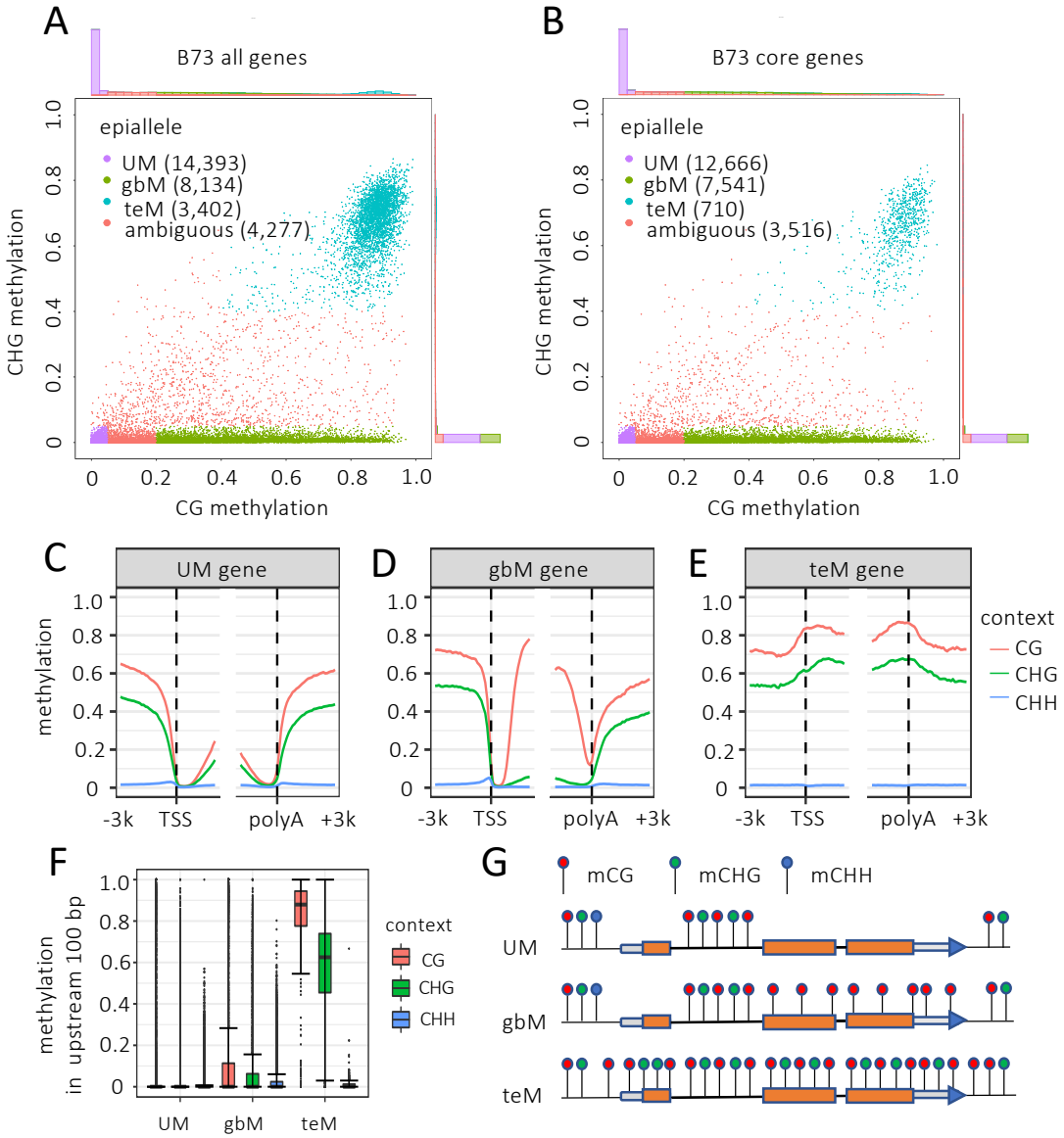


Figure 1

A-B. Scatter plot of mCG vs. mCHG for all B73 genes (A) and core B73 genes (B). Only methylation of coding DNA contributes to methylation values for each gene. Histograms outside axes indicate gene counts in each range of methylation values. Only genes with sufficient coverage of EM-seq reads were included in this analysis (at least 40 cytosines in each context spanned by reads).

C-E. Metagene mCHH , mCHG , and mCG for core UM, gbM, teM B73 genes. Genes are aligned at transcription starts sites (TSSs) and polyadenylation sites (polyA), including 3 Kb upstream, 3 Kb downstream, and 2 KB internal. Methylation values are measured on 100-bp intervals.

F. Distribution of mCHH, mCHG, and mCG for the upstream 100-bp regions for UM, gbM, and teM B73 core genes. Whiskers indicate 1.5 times the interquartile range (IQR).

G. Schematic of gene methylation (epiallele) types. Lollipops indicate methylated cytosines, color coded by context.

0.2 mCG. Both UM and gbM had less than or equal to 0.05 mCHG. We defined genes in the second group as teM genes based on at least 0.4 mCG and 0.4 mCHG. This produced 14,393 UM genes, 8,134 gbM genes and 3,402 teM genes. The remaining 4,277 genes with intermediate methylation values were left uncategorized, along with 9,550 genes with insufficient methylome sequencing read coverage to confidently assign methylation epiallele status.

teM genes are poorly conserved among maize lines

To enrich for functional genes over gene annotation artifacts, we made use of the core gene categorization scheme previously developed by comparison of the 26 NAM founder genomes (Hufford *et al.* 2021). Core genes are the subset that are present at syntenic positions in all 26 genomes, based on intact structural features or RNA expression evidence. These include all copies of tandemly duplicated genes and gene fragments. Core genes are enriched for synteny with Sorghum and for detectable RNA expression relative to the complete set of annotated genes (Hufford *et al.* 2021). 28,289 of 39,756 annotated B73 genes are core genes. Repeating the above gene methylation categorization scheme with only core genes had little effect on the numbers of UM and gbM genes and retained the continuum of mCG in these categories, but it produced a 79.13% decrease in the proportion of teM genes, down from 3,402 to 710 (Fig. 1B). This decrease in the number of teM genes among core genes suggests that many are gene fragments or mis-annotated TEs. For all subsequent analyses, we included only core genes, unless otherwise indicated. While we defined teM genes solely based on methylation in CDS, they also had high CG and CHG methylation at their annotated transcription start sites (TSSs) and polyadenylation sites. This was evident both from metagene methylation profiles (Fig. 1C-E), as well as the

distribution of methylation levels in the first 100 bp upstream of TSSs (Fig. 1 F). Figure 1G provides a simplified summary of methylation profiles for the three gene methylation types.

gbM genes are long and have more intronic TEs

Comparison of structural features of the genes in each methylation category revealed that all components of gbM genes (UTRs, CDSs, introns, and intronic TEs) were longer than UM genes, producing an average 2.9-fold longer total gene lengths (Fig. 2A). Introns had the largest difference in length: the average cumulative intron length of gbM genes was 4,417 bp longer than UM genes (5,718 bp - 1,301 bp). This could not be completely explained by intronic TEs, because the average cumulative TE length in introns was only 1,627 bp longer in gbM than UM genes (2,169 bp - 542 bp). gbM genes were more likely to contain intronic TEs of all superfamilies (Fig. 2B).

teM genes often lack UTRs and have short CDSs that overlap TEs

teM genes were distinguished from both gbM and UM genes by short or absent UTRs (Fig. 2A and C). 59.0% of teM genes lacked both 5' and 3' UTRs, compared to 0.38% of gbM and 5.7% of UM genes. teM genes had relatively short CDSs (816 bp for teM genes, compared to 1,875 bp for gbM and 1,079 bp for UM genes) and tended to overlap annotated TEs: 30.3% of teM genes had at least 100 bp of overlap between CDS and TEs, compared to 5.3% of gbM and 3.4% of UM genes (Fig. 2D).

Narrow expression pattern of UM genes, broad expression of gbM genes, and lack of expression of teM genes

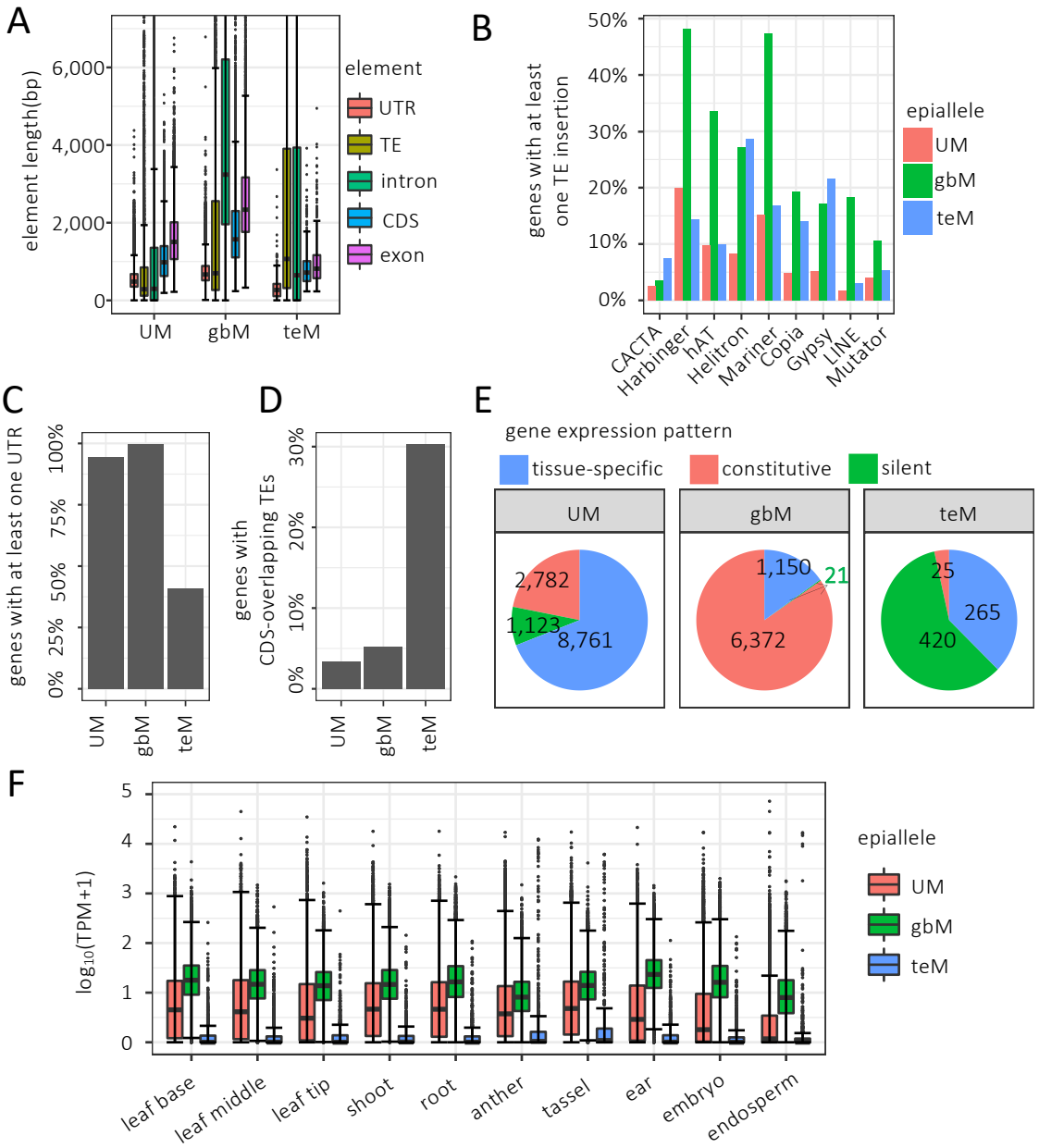


Figure 2

A. Distributions of total CDS length, exon length, intron length, intron TE length, and UTR lengths, in UM, gbM and teM genes, lengths are cumulative, e.g, all exons in a gene are summed to yield a single value for the gene. Whiskers indicate 1.5 IQR.

B. Proportion of genes containing at least one TE insertion in their introns. All nine TE superfamilies are significantly different between UM and gbM (P-value < 0.01, one-tailed Mann-Whitney test).

C. Proportion of genes with at least one annotated UTR.

D. Proportion of genes with at least 100 bp overlap of CDS and annotated TEs.

E. Proportion and number of genes in each gene expression category.

F. Distribution of TPM values for UM, gbM and teM genes across tissue types. Log base 10 of TPM values are shown, and Y-axis is truncated at +5 and -5.

To investigate expression patterns for each category of gene methylation, we used the RNA-seq data from the ten tissues for each NAM founder inbred (Hufford *et al.* 2021). UM genes had larger expression ranges in each tissue than gbM genes, showing both higher and lower extremes in TPM values, consistent with tissue-specific expression (Fig. 2E, F). In contrast, gbM genes were consistently expressed at moderate levels across the ten tissues, consistent with constitutive gene expression (Takuno and Gaut 2012, 2013; Niederhuth *et al.* 2016). teM genes were poorly expressed across all tissues. We categorized genes as tissue-specific based on a TPM value of less than 1 in at least one but not all tissues, as constitutively expressed based on a TPM value of at least 1 in all ten tissues, or as constitutively silent based on a TPM value of less than 1 in all tissues. This expression categorization scheme correlated with methylation categories, where UM genes were most tissue-specific, gbM most constitutive, and teM most silent (Fig. 2E).

Categorizing genes by methylation in NAM founder genome annotations

We applied the same methods used in B73 to categorize genes as UM, gbM or teM to the 25 other NAM founder genomes. Specifically, we used the methylome data previously analyzed, where each set of EM-seq reads was mapped to its own genome and analyzed with respect to its own gene annotations (Hufford *et al.* 2021). For this analysis we included all gene annotations, not just core genes. The abundance of identified UM genes varied from 33% of total genes annotations in M37W to 40% in CML247 (Supplemental Figure 1). The abundance of gbM genes varied from 14% in CML247 to 21% in M37W. The higher abundance of UM genes and lower of gbM genes in CML247 is consistent with the low genome-wide mCG previously observed in CML247 (Hufford *et al.* 2021).

The abundance of teM genes varied from 8.4% in Ky21 to 10% in CML247. Although some teM genes may be demethylated as a means to regulate gene expression, most are likely mis-annotations or pseudogenes. The teM genes are displayed as genome browser tracks for all 26 genomes hosted by the Maize Genetics and Genomics Database (maizeGDB) (Woodhouse *et al.* 2021). They are also listed along with UM and gbM genes for all 26 genomes at https://github.com/dawelab/Natural-methylation-epialleles-correlate-with-gene-expression-in-maize.

Tandem duplication preserves epiallele state

When genes are duplicated, their methylation or expression states may change. To specifically test for epiallele switches associated with tandem duplication, we looked for genes that were present as singletons in at least two genomes but as tandem duplicates in at least one other. For this purpose, we used the pangene system to link homologous syntenic genes (Hufford *et al.* 2021). Each of the 27,910 core pangenes is represented by 26 genes, or in the case of tandem duplicate genes, more than 26 genes because tandem duplicates are linked to a single pangene (Fig. 3A). However, tandem duplications often capture only a fragment of the gene. A single gene can also appear to be a tandem duplicate because 5' and 3' portions are incorrectly annotated as separate genes. To avoid both these issues, we included only intact tandem duplicates for all pangene analyses. We defined intact tandem duplicates as those whose lengths differed by no more than 10% from the median length of all singletons for a given pangene. Since gene dosage effects provide a strong constraint on gene copy number (Birchler and Veitia 2012), we assumed for the purposes of this analysis that singletons best represent ancestral epialleles and tandem duplicates best represent derived ones. 6,270 pangenes exist as tandem

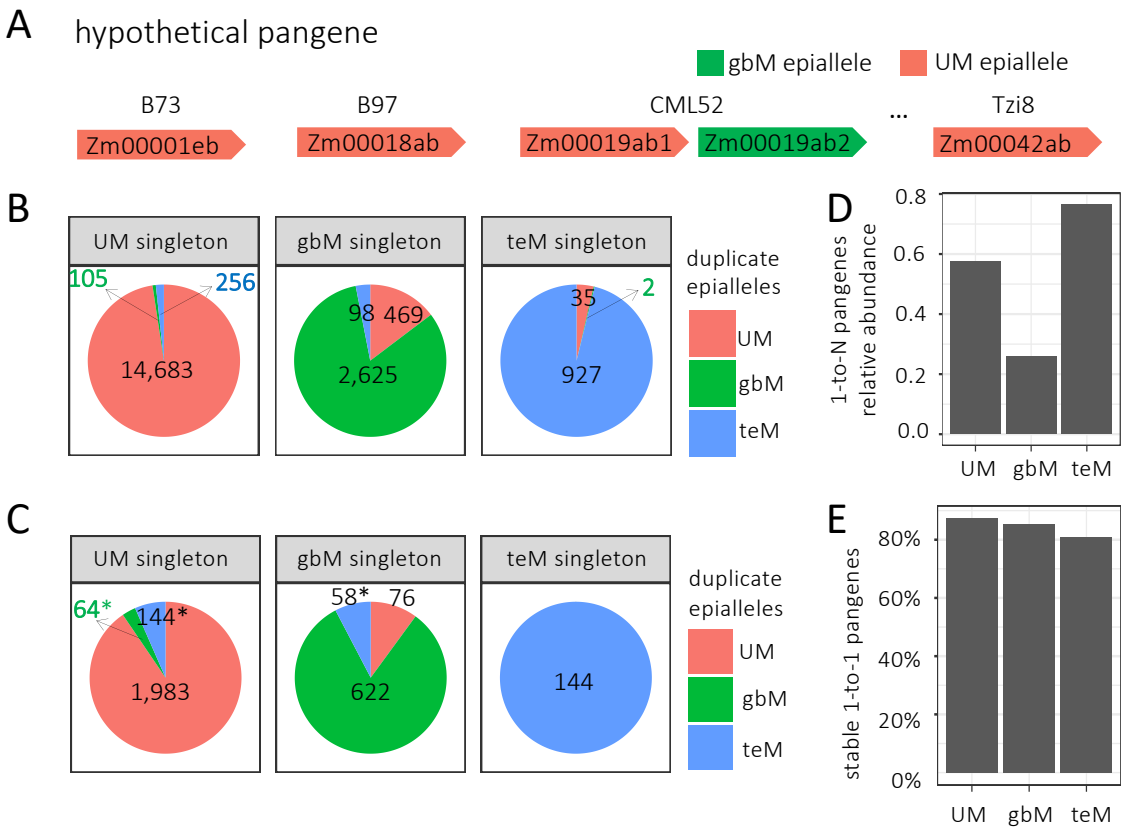


Figure 3

A. Schematic of the relationship between a hypothetical 1-to-N pangene and its individual genes.

B. Proportion of tandem duplication genes of each epiallele type according to corresponding singleton epiallele.

C. Proportion of tandem duplication genes of each epiallele type according to corresponding singleton epiallele, but only including tandem duplication genes with a copy number of at least four. Asterisks indicate a significant difference in proportion from the corresponding value in (B), (p value < 10^{-12}, Chi-square test).

D. Relative abundance of 1-to-N pangenes. 1-to-N pangenes were categorized by their singleton epiallele type, and the total number divided by the number of stable 1-to-1 pangenes of the same epiallele type.

E. Percent of 1-to-1 pangenes with stable epialleles. For each epiallele type, the number of 1-to-1 stable pangenes (represented by just one epiallele) was divided by the number of unstable and stable 1-to-1 pangenes with that epiallele.

duplicates in some inbreds and singletons in at least two other inbreds. Among these 6,270 pangenes, 379 of the singletons were present in more than one epiallele state. We excluded these from further analysis to avoid ambiguity about which epiallele was ancestral. The remaining 5,891 pangenes, represented by singletons of one epiallele type in at least two genomes, were represented by arrays of up to 42 duplicates in other inbreds. We referred to these as "1-to-N" pangenes.

Comparing the singleton epialleles with their corresponding duplicate epialleles in other genomes revealed that epiallele states of the singletons were normally maintained in the duplicates (Fig. 3B). In the case of UM singletons, 98% of duplicates were also UM. This is a surprisingly high percent because a change in mCG value as small as 0.15 is enough to switch between UM and gbM. A larger change of 0.35 in both mCG and mCHG is required for changing between UM and teM, yet switches of this type were more common than UM to gbM (1.7% of duplicates vs 0.7% of duplicates). Switches to teM were even more common (6.6%) in the subset of duplicates with a minimum copy number of four (Fig. 3C). A similar trend was present in duplicates from gbM singletons: teM duplicates made up 3% of duplicates in the complete set of duplicates, but 7.6% in the minimum-four-copy set. These increases in abundance of teM duplicates in the minimum-four-copy duplicates were significant for both UM and gbM singletons (P-value $< 10^{-10}$, Chi-square test).

gbM genes are poorly represented among tandem duplications

Of the 5,891 1-to-N pangenes, 4,063 had UM singleton epialleles, 1,538 had gbM singleton epialleles, and 290 had teM singleton epialleles. To put these numbers in perspective, we compared the numbers to stable 1-to-1 pangenes, which had only singletons and only one

epiallele type. As with 1-to-N pangenes, partial duplications were ignored in determining singleton vs. duplicate status. After excluding pangenes that had fewer than two genes with defined epialleles, 7,061 of the stable 1-to-1 pangenes had only UM epialleles, 5,953 had only gbM, and 379 had only teM. Normalizing the numbers of 1-to-N pangenes by the numbers of stable 1-to-1 pangenes suggests that UM singletons are 2.3-fold more likely to be associated with gene duplications than gbM singletons (Fig. 3D). gbM epialleles may be less susceptible to tandem duplication or have more severe fitness consequences than duplicates derived from UM epialleles.

Stability of epialleles in 1-to-1 pangenes

An additional 1,056 pangenes occurred as singletons in all genomes but were present in more than one epiallele state. We called these unstable 1-to-1 pangenes. There were 966 UM-gbM switches, 43 UM-teM switches, 32 gbM-teM switches, and 15 UM-gbM-teM switches. To be included in one of these unstable 1-to-1 pangene groups, we required at least two genomes contain each epiallele. The remaining 7,570 core pangenes did not meet the stringent requirements for 1-to-N, stable 1-to-1, or unstable 1-to-1 pangenes. The dominance of UM-gbM switches reflects the larger number of UM and gbM epialleles. teM epialleles, though rare, are more likely to be unstable than either UM or gbM. A total of 13% of the UM, 15% of the gbM, and 19% of the teM epiallele-containing pangenes (Fig. 3E) were unstable in the NAM founder inbreds (calculated as unstable pangenes divided by stable plus unstable pangenes).

gbM epialleles are expressed higher than UM epialleles

Maize is a pseudotetraploid where the majority of genes have two copies that often have redundant functions (Woodhouse *et al.* 2010). Thus maize offers an opportunity to test relationships between epialleles and gene expression changes that would not be tolerated in plants where most genes are single copy. For each of the 966 unstable 1-to-1 UM-gbM pangenes, we calculated the differences in the mean transcripts per million (TPM) for UM epialleles and gbM epialleles, which we refer to as the gbM-UM TPM differences. Epialleles with large TPM values create a broad and noisy distribution of gbM-UM TPM differences. Nonetheless, the distribution would be expected to center on a value of zero if UM-gbM epiallele switches are not associated with gene expression change. We found that the median gbM-UM TPM difference was above zero for all ten tissues (Figure 4A and B). These differences were significantly different (binomial sign test P-value $< 10^{-10}$, for all ten tissues and Wilcoxon signed rank test P-value ranging from 0.02 to 1.7×10^{-4}). Normalizing the median gbM-UM TPM differences by mean UM TPM values for the same set of pangenes in each tissue produced highly consistent results. The median gbM-UM TPM differences indicate that gbM epialleles are expressed about 3% higher than UM epialleles. The differences were lowest in root (at 1.6%) and highest in leaf tip (at 3.7%). We used median values for these analyses because the means are skewed by epialleles with large TPM values, where very minor changes in expression manifest as very large changes in TPM value. For example, a 10% change in expression of a gene with a TPM value of 1000 would give a gbM-UM TPM difference of 100 TPM. Even so, all tissues but root yielded mean gbM-UM TPM difference of greater than zero.

As comparisons, we also examined epiallele expression in the 43 UM-teM and 32 gbM-teM unstable 1-to-1 pangenes. The small numbers of these pangenes preclude meaningful

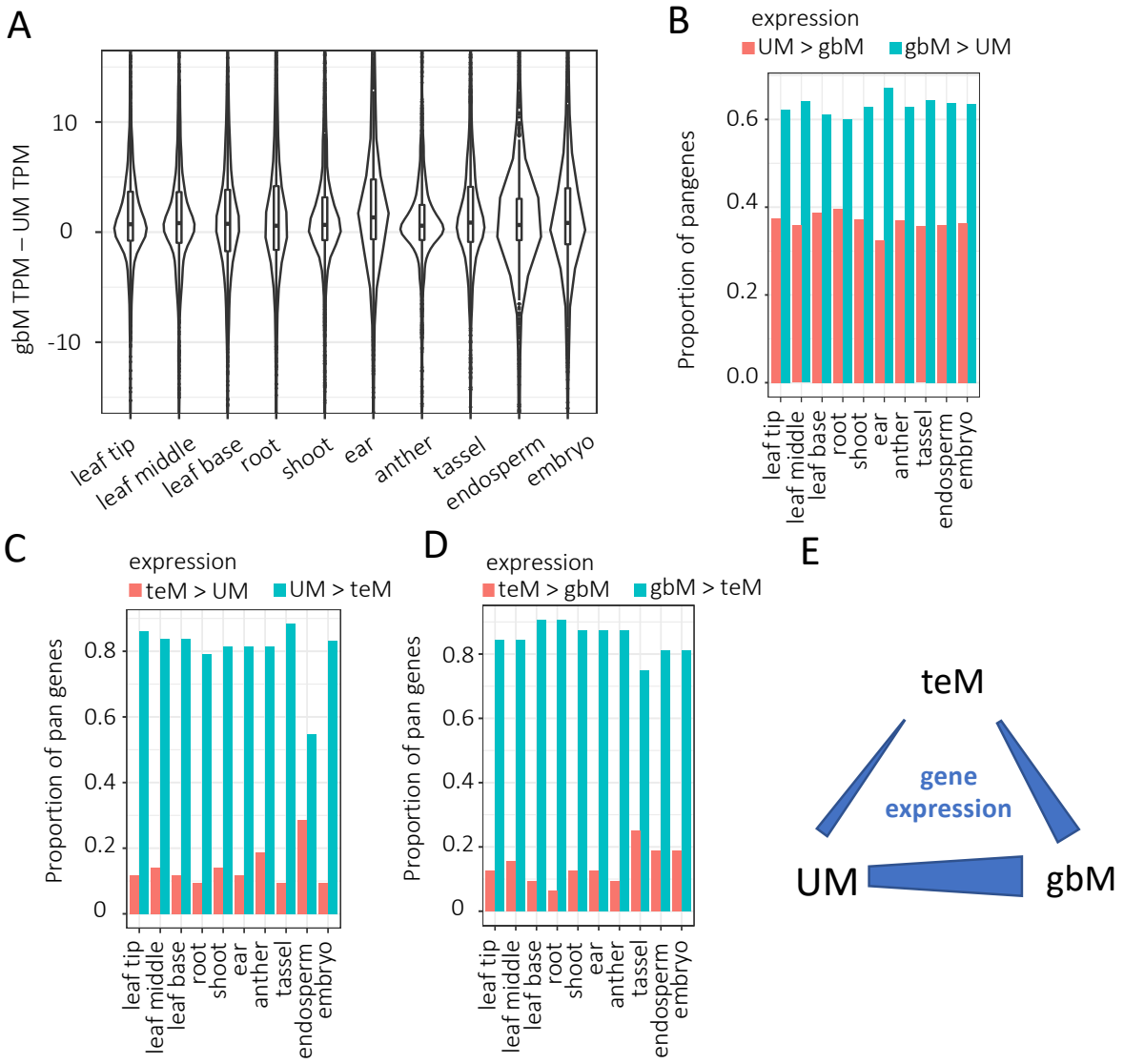


Figure 4

A. Distribution of differences in TPM between gbM and UM epialleles for each 1-to-1 unstable pangene.

B. Proportion of 1-to-1 unstable pangenes with differences in TPM between gbM and UM epialles that were either greater or less than zero.

C. Proportion of 1-to-1 unstable pangenes with differences in TPM between UM and teM epialles that were either greater or less than zero.

D. Proportion of 1-to-1 unstable pangenes with differences in TPM between gbM and teM epialles that were either greater or less than zero.

E. Schematic of relationship between epialleles and gene expression.

quantification of expression changes for teM epialleles, but there was a clear trend for teM epialleles to have reduced expression relative to UM or gbM (Fig. 4C and D).

DISCUSSION

Gene body methylation defines a continuum of genes, while TE-like methylation defines a distinct group that is enriched for annotation artifacts.

Examining CG methylation alone reveals a clear bimodal distribution, with most genes having either near-zero CG methylation or greater than 70% methylation (Fig.1A and B). After removing the set of genes with high CHG methylation, which is characteristic of heterochromatin, the remaining genes show a range of CG methylation heavily weighted toward zero CG methylation. Thus a categorization of non-CHG methylated genes as either unmethylated (UM) or gene body methylated (gbM) is a simplification of a continuously distributed feature. Despite the limitations of simplified UM-gbM binary categorization schemes, they do correlate with structural and expression features and reveal hints of both the origins and consequences of gene body methylation [Reviewed in (Bewick and Schmitz 2017; Muyle *et al.* 2022)].

Examining both CG and CHG methylation together, however, reveals a distinct TE-like methylation (teM) category. We found multiple lines of evidence that this category is dominated by nonfunctional genes and annotation artifacts. First, they are poorly conserved among maize lines (Fig. 1A, B). Second, their methylation extends into their 5' and 3' flanking sequences, indicating a lack of functional cis regulatory sequence (Fig. 1C, D). Third, they frequently lack both 5' and 3' UTRs, have short CDSs, and their CDSs overlap with annotated TEs (Fig. 2A, C, D). Fourth, they are poorly expressed (Fig. 2E, F and 4C, D). Fifth, teM epialleles are rare

among core pangenes (genes present in all 26 NAM lines), and when present they often have UM or gbM epialleles as well (Fig 3E). These data support the view that most teM genes are usually inactive or nonfunctional. They are likely pseudogenes, fragmented tandem duplications, TEs, or annotation errors. TE-like methylation alone, however, should not completely exclude the possibility of gene function, as it remains possible that some functional genes may utilize TE-like methylation as a means of developmental gene regulation.

Addition of a "TE-like methylation" flag to gene annotations

Rather than removing teM genes from annotations, we suggest flagging them as having TE-like methylation. Based on this information, users may choose to give teM genes lower priority when identifying gene candidates for functional studies. Annotation tracks showing teM genes from our study are now available on the NAM founder genome browsers hosted by maizeGDB and make this information easily accessible.

Gene body methylation is associated with stable gene copy number

We found that gbM genes were strongly underrepresented among genes that give rise to intact tandem duplications (Fig. 3D). Since these genes tend to have stable expression over a broad range of cell types (Fig. 2E, F), we speculate that increase in gene dosage of gbM genes tends to be more disruptive than that of UM genes. This would be consistent with prior work showing stronger genetic conservation of gbM genes across species (Takuno and Gaut 2013; Niederhuth *et al.* 2016; Seymour and Gaut 2020).

Gene body methylation epialleles are associated with weak but significant increases in expression

The maize NAM founder genomes, methylomes, and transcriptomes provide a powerful resource for studying relationships between methylation and gene expression. The quality of the genome assemblies and associated annotation allowed us to filter out genes whose CDS had diverged too much for useful comparisons, accurately call methylation epiallele states in each genome; and accurately access gene expression levels in each genome. We found that gbM epialleles correlate with expression increases relative to UM epialleles in ten different tissues (Fig. 4A,B), consistent with expression data from *E. salsugineum* and *A. thaliana* mutants showing that a loss of gene body methylation can lead to reduced gene expression (Shahzad *et al.*; Muyle *et al.* 2021) The fact that the expression changes were small, with gbM genes showing ~3% higher expression than UM genes, is consistent with the fact that gene body methylation is dispensable in some plants (Bewick *et al.* 2016). Maize, with its relatively high levels of genetic redundancy (Woodhouse *et al.* 2010) may be more tolerant of epiallele switches than species such as Arabidopsis with more streamlined genomes.

On the theoretical importance of gene body methylation

Our results support recent data suggesting that gbM is correlated with subtle increases in gene expression (Shahzad *et al.*; Muyle *et al.* 2021), although we note with caution that the validity of earlier results supporting similar conclusions (Muyle and Gaut 2019) have been questioned (Bewick *et al.* 2019). It is also unclear whether changes in gene body methylation are a cause of changes in gene expression or a consequence of other events that also affect gene expression. One possibility for how gbM could affect transcription is that it prevents internal

transcription initiation that would interfere with normal transcription, as has been reported in Arabidopsis and in animals (Neri *et al.* 2017; Teissandier and Bourc'his 2017; Choi *et al.* 2020). However DNA methylation and other linked chromatin modifications affect more than transcription. In theory, any process that involves enzymes making contact with DNA could be inhibited or facilitated (e.g., DNA repair, recombination, replication, transposon integration). Further, functions associated with gbM may be important in some species but dispensable in others.

METHODS

Categorizing genes by methylation epiallele and metagene methylation analysis

CGmap files produced using BS-Seeker2 software in the NAM founder study were used as the input for all DNA methylomes analyses (Guo *et al.* 2013; Hufford *et al.* 2021). Each methylome was analyzed relative to its own reference genome, as opposed to the simpler but less accurate method of using B73 as the reference for all. The CGmapTools mtr tool was used to calculate average methylation values of each gene using the "by region" method after filtering the CGmaps specifically for CDS using the CGmaptools select region tool (Guo *et al.* 2013, 2018). Gene annotations were obtained from https://download.maizegdb.org. Only canonical gene annotations were used in defining CDS, as well as for all other genic features. Only genes with at least 40 cytosines in the CG context and 40 cytosines in the CHG context spanned by EM-seq reads were assigned methylation epialleles. UM epialleles were defined by both mCG and mCHG less than 0.05, gbM epialleles by mCG higher than 0.2 and mCHG less than 0.05, and teM epialleles by both mCG and mCHG methylation levels higher than 0.4. The metagene methylation values over 3 Kb upstream and downstream of genes and 1.5 Kb within genes were

produced using the CGmapTools mfg tools with 100 bp intervals and minimum coverage of 1 (-c 1 parameter). First, however, the CGmapTools bed2fragreg tool was used in combination with an awk command to convert input gene annotations from BED format to the fragreg format used as input for the mfg tool.

Identification of core genes and quantification of genic structural features

To identify core genes, which were annotated as genes in B73 and in all other 25 NAM founders, we made use of a pangene table that lists every gene in all 26 genomes with each row corresponding to a single pangene and each column as a single NAM founder. The pangene table was downloaded from https://de.cyverse.org/anon-files//iplant/home/shared/NAM/NAM_genome_and_annotation_Jan2021_release/SUPPLEMENTAL_DATA/pangene-files/pan_gene_matrix_v3_cyverse.csv (Hufford *et al.* 2021). Tandem duplicate genes are included as multiple genes within a single cell. "NA" indicates a missing gene. Genome coordinates instead of gene names indicate presence of DNA homologous to the pangene but insufficient evidence for a gene annotation. These were not included in our analyses because they lack gene annotations. To calculate the length of genic structural features, we subtracted end from start coordinates using awk commands on annotation files, and we summed the individual lengths per gene using the R aggregate function. Intronic repeat lengths were obtained using the BEDTools v.2.30 (Quinlan and Hall 2010) intersect tool with -wo -wa -a (introns) -b (repeats). Repeat annotations are from https://download.maizegdb.org. The intersected repeat annotations were merged using the BEDTools merge tool to prevent overlapping repeat annotations from being counted twice.

To calculate numbers of genes with TE insertions, we used the Python pandas package (Python-3.7.4 environment) to generate simplified bed files for TE superfamilies. The BEDTools intersect tool with -wa -wb -a (intron) -b (TE) was used to identify TEs in introns. The R unique function was used to count each gene only once even if multiple annotations of the same TE superfamily were present. The R merge function was used to associate each gene with an insertion with its methylation epiallele status. The R table function was then used to count numbers of genes with at least one insertion for each superfamily.

To calculate extents of overlap between CDS and repetitive elements, we used the BEDTools intersect tool with -wo -a (CDS) -b (repeats). The R aggregate function was used to sum the CDS-overlapping repeat lengths per gene. The R merge function was used to associate each gene with its methylation epiallele status. "NA" was replaced with 0 for genes with no overlap between CDSs and repeats. The R table function was then used to count numbers of genes with CDS-overlapping repeat lengths that were greater than 100 bp.

Calculating gene TPM values and assigning expression categories

To quantify gene expression, we mapped mRNA-seq reads from the NAM assembly project using similar methods to the NAM assembly project (Hufford *et al.* 2021). Briefly, STAR software (version 2.7.2) (Dobin *et al.* 2013) was used to map reads to each of the 26 genomes assemblies and their reference gene annotations. Unlike the NAM assembly project, however, gene annotations were used to guide read mapping with the --sjdbGTFfile and --twopassMode Basic parameters. Prior to read mapping, Cufflinks software (version 2.2.1) (Trapnell *et al.* 2010) was used to convert gff3 gene annotations into gft format. As in the NAM assembly project, transcripts per million value (TPM), were obtained based on read counts per genes from

featureCounts software (Liao *et al.* 2014) using uniquely mapping reads only (default method). For tissues with two mRNA-seq replicates, both replicates were merged.

To define the gene expression categories, we compared TPM values across the ten tissues (leaf tip, leaf middle, leaf base, root, shoot, ear, anther, tassel, endosperm, and embryo). We defined tissue-specific expression as TPM >= 1 in at least one tissue and TPM < 1 in at least one tissue, constitutive expression as TPM >= 1 in all ten tissues, and silent as TPM < 1 in all ten tissues. TPM values from all ten tissues were combined into one matrix, and the matrixStats package in R was used to identify each expression category.

Conceptual summary of pangenes analysis

A schematic summary of our pangene methods is shown in Figure 5. In brief, the three major input data sources—DNA methylomes, RNA transcriptomes, and gene annotations–were used to create a series of gene matrices, where the row coordinate indicates the pangene and the column coordinate indicates the NAM founder. The matrices included values such as CDS length, TPM, and epiallele status. Tandem duplicate genes were represented by lists of values within single cells. The initial matrices were intersected with each other using specific criteria such as CDS length to produce filtered matrices representing subsets of genes of particular interest. Values in these filtered matrices were then extracted and used as inputs for plotting distributions (such as TPM) or numbers of genes with specific features like stable epialleles.

Identification of genes with intact CDS

To exclude both annotation artifacts and genes with large structural changes in their CDSs, we required that the CDS length vary by less than 10% from the median of each pangene.

Figure 5. Schematic of pangene analysis workflow

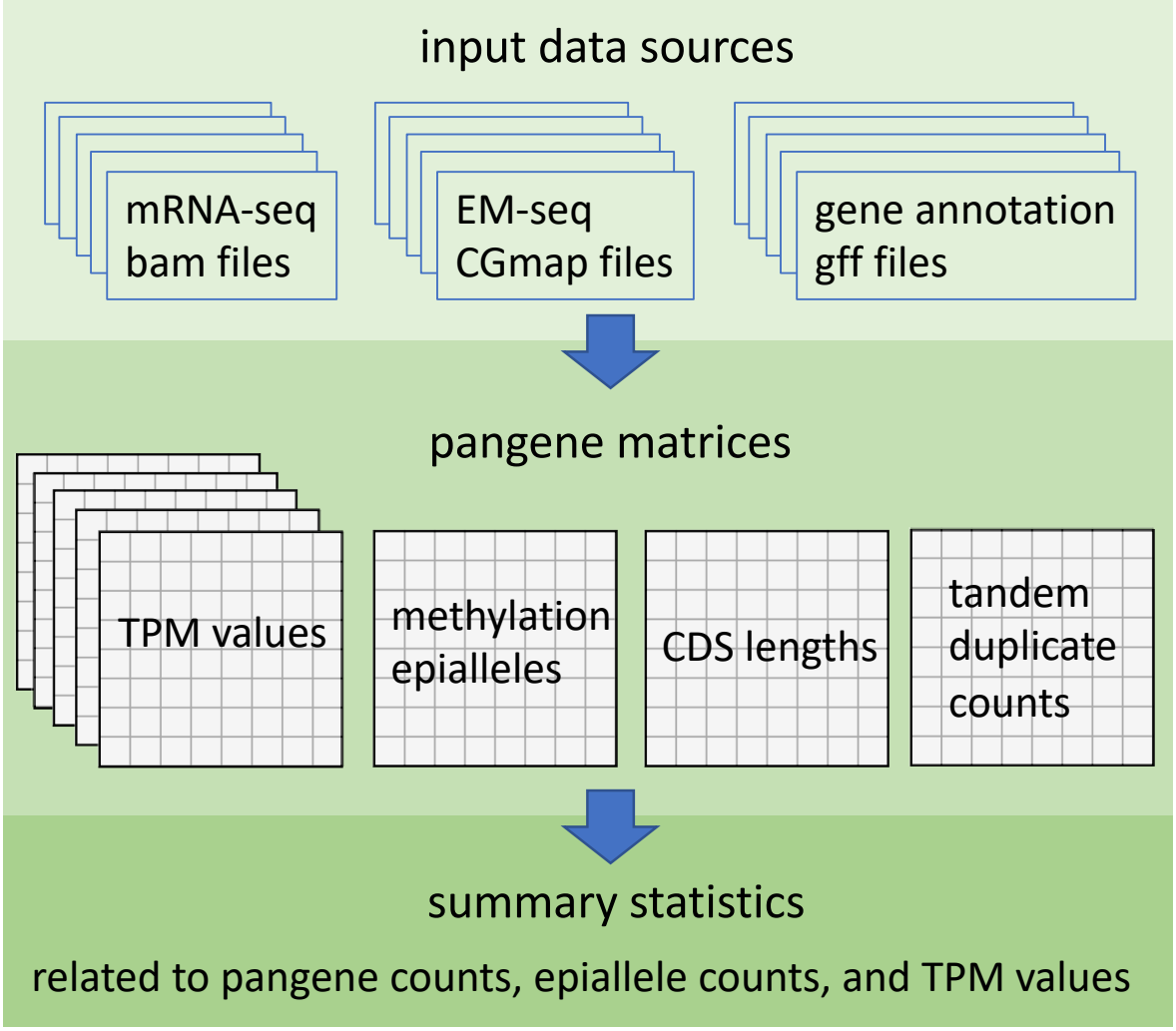


Figure 5
Input data sources for each NAM founder genome were either downloaded from maizeGDB or produced according to methods of the NAM assembly project (Hufford et al 2021). Pangene matrices consisted of one pangene per row. Each column corresponded to a NAM founder, with values from either single genes or from clusters of tandem duplicate genes.

Only singletons were included in determining the median length. Since CDS lengths for tandem duplicates were represented by lists of values within individual cells, the R rowMedians function was used to determine median lengths (rowMedians ignores values in lists). Both singletons and tandem duplicates were compared to the same singleton-defined median length and unqualified genes were removed using R logical operators. Removing genes by this criterion did not affect their core gene status, only whether they were included in subsequent analyses. Input gene sets for all pangene analyses were limited by this CDS length restriction.

Identification of 1-to-1 and 1-to-N pangenes and counting epialleles

To identify pangenes with and without tandem duplicates, we created a pangene matrix of tandem duplicate counts, where a value of one indicated a singleton, a value of two indicated two copies, etc. The R rowMins function was used to identify a set of pangenes with both singletons and duplicates. The R rowSums function was used to count numbers of singleton genes of each defined epiallele type (UM, gbM and teM) for each pangene. Pangenes that had more than one singleton epiallele type or had only one singleton gene with a defined epiallele were removed to produce the final 1-to-N pangene matrix. To count epialleles among tandem duplicates, the singleton epialleles (any cells not containing lists) were first converted to nulls. Then all cells containing lists were combined with the R unlist function and the epialleles counted with the R table function.

The stable 1-to-1 pangenes matrix was produced similarly as the 1-to-N pangenes, except it was derived from singleton-only pangenes. The unstable 1-to-1 pangene matrix was also derived from singleton-only pangenes, but only included pangenes represented by at least two

epiallele types, where each epiallele type was represented by at least two genes. This produced four types of unstable 1-to-1 pangenes: UM-gbM, UM-teM, gbM-teM, and UM-gbM-teM.

Comparison of epiallele expression levels in unstable pangenes

To compare epiallele expression values among unstable 1-to-1 pangenes, we calculated the mean TPM value for each epiallele type in each pangene individually using the following formula:

$$trace\{DT_{ij} \times I^T(DE_{ij})\} \oslash trace\{\mathbf{1}_{ij} \times I^T(DE_{ij})\}$$

Variables are as follows: i is the pangene index and j is the genome index. DT_{ij} (tissue) is a matrix of TPM values, one for each tissue. DE_{ij} is a matrix of epialleles. $I(DE_{ij})$ is an indicator matrix of zeros and ones, one indicator matrix for each of the three epiallele types. DT_{ij} (tissue) times the transpose of $I(DE_{ij})$ and divide sum of the transpose of $I(DE_{ij})$ over j was used to create three pangene lists for each tissue, where each list contains the mean TPM value for one 1epiallele type for that pangene. These lists were then used to calculate TPM differences for each epiallele type. To measure mean expression values of epiallele types in unstable 1-to-1 pangenes across the whole set, the R unlist function was used to make lists of TPM values for each epiallele type using the TPM matrices and epiallele matrix as inputs. The R summary function was used to calculate summary statistics from these lists.

P-value calculations

To test whether the proportions of epiallele types differed between tandem duplicates of 1-to-N pangenes with minimum of two copies and 1-to-N pangenes with a minimum of four

copies in Figure 3B and C, we applied Chi-square tests on epiallele counts using the R function chisq.test (x,y,correct=F). To test whether the direction of TPM differences between epiallele types was significantly different (gain or loss of expression in Figure 4B-D), we applied both binomial sign tests and Wilcoxon signed rank tests on each set of pangene counts. Binomial sign tests were done using the R function binom.test(sum(gbM>UM),n,p=0.5,alternative = "two.sided") and the Wilcoxon signed rank test using the R function wilcox.test().

Github link for scripts and R methods used in this study

https://github.com/dawelab/Natural-methylation-epialleles-correlate-with-gene-expression-in-maize

COMPETING INTERESTS

The authors declare that they have no competing interests.

References

Anderson S. N., and N. M. Springer, 2018 Potential roles for transposable elements in creating imprinted expression. Curr. Opin. Genet. Dev. 49: 8–14.

Baubec T., D. F. Colombo, C. Wirbelauer, J. Schmidt, L. Burger, *et al.*, 2015 Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. Nature 520: 243–247.

Bewick A. J., L. Ji, C. E. Niederhuth, E.-M. Willing, B. T. Hofmeister, *et al.*, 2016 On the origin and evolutionary consequences of gene body DNA methylation. Proc. Natl. Acad. Sci. U. S. A. 113: 9111–9116.

Bewick A. J., and R. J. Schmitz, 2017 Gene body DNA methylation in plants. Curr. Opin. Plant Biol. 36: 103–110.

Bewick A. J., Y. Zhang, J. M. Wendte, X. Zhang, and R. J. Schmitz, 2019 Evolutionary and Experimental Loss of Gene Body Methylation and Its Consequence to Gene Expression. G3 9: 2441–2445.

Birchler J. A., and R. A. Veitia, 2012 Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. Proc. Natl. Acad. Sci. U. S. A. 109: 14746–14753.

Borg M., R. K. Papareddy, R. Dombey, E. Axelsson, M. D. Nodine, *et al.*, 2021 Epigenetic reprogramming rewires transcription during the alternation of generations in Arabidopsis. Elife 10. https://doi.org/10.7554/eLife.61894

Bröhm A., T. Schoch, M. Dukatz, N. Graf, F. Dorscht, *et al.*, 2022 Methylation of recombinant mononucleosomes by DNMT3A demonstrates efficient linker DNA methylation and a role of H3K36me3. Commun Biol 5: 192.

Choi J., D. B. Lyons, M. Yvonne Kim, J. D. Moore, and D. Zilberman, 2020 DNA Methylation and Histone H1 Jointly Repress Transposable Elements and Aberrant Intragenic Transcripts. Molecular Cell 77: 310–323.e7.

Crisp P. A., A. P. Marand, J. M. Noshay, P. Zhou, Z. Lu, *et al.*, 2020 Stable unmethylated DNA demarcates expressed genes and their cis-regulatory space in plant genomes. Proc. Natl. Acad. Sci. U. S. A. 117: 23991–24000.

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

Drinnenberg I. A., D. deYoung, S. Henikoff, and H. S. Malik, 2014 Recurrent loss of CenH3 is associated with independent transitions to holocentricity in insects. Elife 3. https://doi.org/10.7554/eLife.03676

Guo W., P. Fiziev, W. Yan, S. Cokus, X. Sun, *et al.*, 2013 BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. BMC Genomics 14: 774.

Guo W., P. Zhu, M. Pellegrini, M. Q. Zhang, X. Wang, *et al.*, 2018 CGmapTools improves the precision of heterozygous SNV calls and supports allele-specific methylation detection and visualization in bisulfite-sequencing data. Bioinformatics 34: 381–387.

Halter T., J. Wang, D. Amesefe, E. Lastrucci, M. Charvin, *et al.*, 2021 The Arabidopsis active demethylase ROS1 cis-regulates defence genes by erasing DNA methylation at promoter-regulatory regions. Elife 10. https://doi.org/10.7554/eLife.62994

Hufford M. B., A. S. Seetharam, M. R. Woodhouse, K. M. Chougule, S. Ou, *et al.*, 2021 De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. Science 373: 655–662.

Kawakatsu T., S.-S. C. Huang, F. Jupe, E. Sasaki, R. J. Schmitz, *et al.*, 2016 Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. Cell 166: 492–505.

Khouider S., F. Borges, C. LeBlanc, A. Ungru, A. Schnittger, *et al.*, 2021 Male fertility in Arabidopsis requires active DNA demethylation of genes that control pollen tube function. Nat. Commun. 12: 410.

Kim J.-S., J. Y. Lim, H. Shin, B.-G. Kim, S.-D. Yoo, *et al.*, 2019 ROS1-Dependent DNA Demethylation Is Required for ABA-Inducible NIC3 Expression. Plant Physiol. 179: 1810–1821.

Le N. T., Y. Harukawa, S. Miura, D. Boer, A. Kawabe, *et al.*, 2020 Epigenetic regulation of spurious transcription initiation in Arabidopsis. Nat. Commun. 11: 3224.

Liao Y., G. K. Smyth, and W. Shi, 2014 featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30: 923–930.

Lyons D. B., A. Briffa, S. He, J. Choi, E. Hollwey, *et al.*, 2022 Extensive de novo activity stabilizes epigenetic inheritance of CG methylation in Arabidopsis transposons. bioRxiv 2022.04.19.488736.

Martin, Seymour, and Gaut, 2021 CHH methylation islands: a nonconserved feature of grass genomes that is positively associated with transposable elements but negatively associated with gene …. Genome Biol. Evol.

Muyle A., and B. S. Gaut, 2019 Loss of Gene Body Methylation in Eutrema salsugineum Is Associated with Reduced Gene Expression. Mol. Biol. Evol. 36: 155–158.

Muyle A., J. Ross-Ibarra, D. K. Seymour, and B. S. Gaut, 2021 Gene body methylation is under selection in Arabidopsis thaliana. Genetics 218. https://doi.org/10.1093/genetics/iyab061

Muyle A. M., D. K. Seymour, Y. Lv, B. Huettel, and B. S. Gaut, 2022 Gene Body Methylation in Plants: Mechanisms, Functions, and Important Implications for Understanding Evolutionary Processes. Genome Biol. Evol. 14. https://doi.org/10.1093/gbe/evac038

Neri F., S. Rapelli, A. Krepelova, D. Incarnato, C. Parlato, *et al.*, 2017 Intragenic DNA methylation prevents spurious transcription initiation. Nature 543: 72–77.

Niederhuth C. E., A. J. Bewick, L. Ji, M. S. Alabady, K. D. Kim, *et al.*, 2016 Widespread natural variation of DNA methylation within angiosperms. Genome Biol. 17: 194.

Ossowski S., K. Schneeberger, J. I. Lucas-Lledó, N. Warthmann, R. M. Clark, *et al.*, 2010 The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. Science 327: 92–94.

Papareddy R. K., K. Páldi, A. D. Smolka, P. Hüther, C. Becker, *et al.*, 2021 Repression of CHROMOMETHYLASE 3 prevents epigenetic collateral damage in Arabidopsis. Elife 10. https://doi.org/10.7554/eLife.69396

Quinlan A. R., and I. M. Hall, 2010 BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842.

Reinders J., B. B. H. Wulff, M. Mirouze, A. Marí-Ordóñez, M. Dapp, *et al.*, 2009 Compromised stability of DNA methylation and transposon immobilization in mosaic Arabidopsis epigenomes. Genes Dev. 23: 939–950.

Schmitz R. J., Z. A. Lewis, and M. G. Goll, 2019 DNA Methylation: Shared and Divergent Features across Eukaryotes. Trends Genet. 35: 818–827.

Seymour D. K., and B. S. Gaut, 2020 Phylogenetic Shifts in Gene Body Methylation Correlate with Gene Expression and Reflect Trait Conservation. Mol. Biol. Evol. 37: 31–43.

Shahzad Z., J. D. Moore, J. Choi, and D. Zilberman, Epigenetic inheritance mediates phenotypic diversity in natural populations. https://www.biorxiv.org/content/10.1101/2021.03.15.435374v2

Shayevitch R., D. Askayo, I. Keydar, and G. Ast, 2018 The importance of DNA methylation of exons on alternative splicing. RNA 24: 1351–1362.

Stroud H., M. V. C. Greenberg, S. Feng, Y. V. Bernatavichute, and S. E. Jacobsen, 2013 Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome. Cell 152: 352–364.

Takuno S., and B. S. Gaut, 2012 Body-methylated genes in Arabidopsis thaliana are functionally important and evolve slowly. Mol. Biol. Evol. 29: 219–227.

Takuno S., and B. S. Gaut, 2013 Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. Proc. Natl. Acad. Sci. U. S. A. 110: 1797–1802.

Teissandier A., and D. Bourc'his, 2017 Gene body DNA methylation conspires with H3K36me3 to preclude aberrant transcription. EMBO J. 36: 1471–1473.

Trapnell C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, *et al.*, 2010 Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28: 511–515.

Wendte J. M., Y. Zhang, L. Ji, X. Shi, R. R. Hazarika, *et al.*, 2019 Epimutations are associated with CHROMOMETHYLASE 3-induced de novo DNA methylation. Elife 8. https://doi.org/10.7554/eLife.47891

Wollmann H., H. Stroud, R. Yelagandula, Y. Tarutani, D. Jiang, *et al.*, 2017 The histone H3 variant H3.3 regulates gene body DNA methylation in Arabidopsis thaliana. Genome Biol. 18: 94.

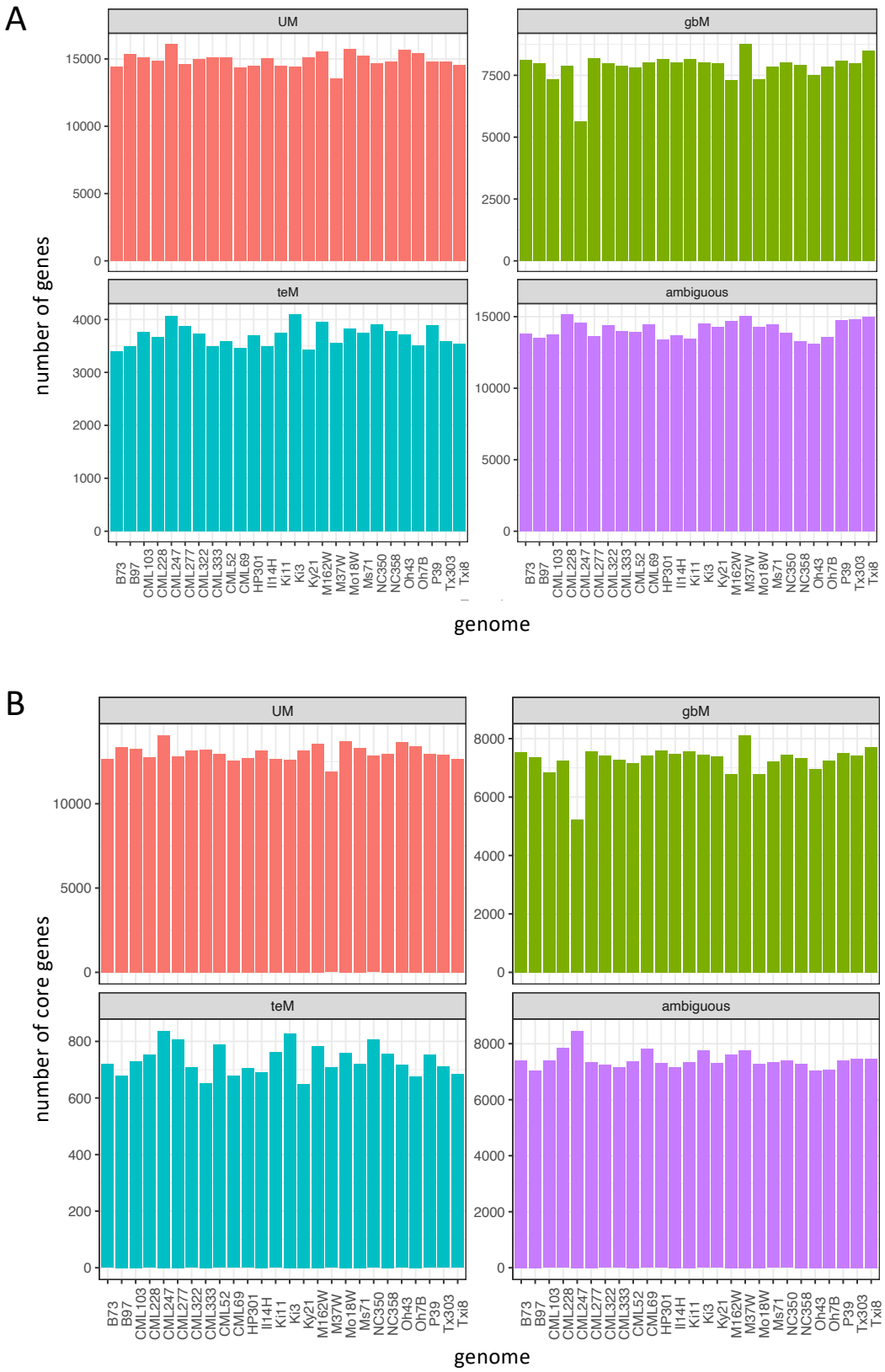
Woodhouse M. R., J. C. Schnable, B. S. Pedersen, E. Lyons, D. Lisch, *et al.*, 2010 Following Tetraploidy in Maize, a Short Deletion Mechanism Removed Genes Preferentially from One of the Two Homeologs. PLoS Biol. 8: e1000409.

Woodhouse M. R., E. K. Cannon, J. L. Portwood 2nd, L. C. Harper, J. M. Gardiner, *et al.*, 2021 A pan-genomic approach to genome databases using maize as a model system. BMC Plant Biol. 21: 385.

Yearim A., S. Gelfman, R. Shayevitch, S. Melcer, O. Glaich, *et al.*, 2015 HP1 is involved in regulating the global impact of DNA methylation on alternative splicing. Cell Rep. 10: 1122–1134.

Zhang Y., J. M. Wendte, L. Ji, and R. J. Schmitz, 2020 Natural variation in DNA methylation homeostasis and the emergence of epialleles. Proc. Natl. Acad. Sci. U. S. A. 117: 4874–4884.

Zhang Y., H. Jang, R. Xiao, I. Kakoulidou, R. S. Piecyk, *et al.*, 2021 Heterochromatin is a quantitative trait associated with spontaneous epiallele formation. Nat. Commun. 12: 6958.



Supplemental Figure 1

Number of genes of each epiallele type in of the 26 NAM founder genomes.

A. All annotated genes.

B. Core genes only.