DNA methylation signatures of duplicate gene evolution in angiosperms

Sunil K. Kenchanmane Raju[1], S. Marshall Ledford[2], Chad E. Niederhuth[1,3,*]

[1]Department of Plant Biology, Michigan State University, East Lansing, MI 48824, U.S.A.

[2]Vassar College, Poughkeepsie, NY 12604, U.S.A.

[3]AgBioResearch, Michigan State University, East Lansing, MI 48824, U.S.A.

[*]Correspondence: Chad E. Niederhuth (niederhu@msu.edu)

ABSTRACT

Gene duplication is a major source of genetic variation and functional diversity. Chromatin modifications have been proposed to have a significant role in their evolution. We analyzed DNA methylation of 43 angiosperm species to identify relationships between DNA methylation and gene duplication. Whole-genome duplicates (WGD) were enriched for gene-body methylated (gbM; associated with broad expression) and unmethylated (unM; associated with narrower expression) genes, and depleted for transposon-like methylated (teM; associated with silencing) genes. Small-scale or single gene duplicates (SGDs) showed different enrichment patterns. Tandem and proximal duplicates were predominantly enriched for unM and variably for teM, while translocated and dispersed duplicates were enriched for teM and depleted in gbM. DNA methylation patterns of duplicate genes were significantly associated with the age of duplication. Younger WGDs tend to be gbM, while older WGDs are often unM. Recently duplicated single-gene duplicates (SGDs) were enriched in teM while gbM was seen in older SGDs. Across 928 *Arabidopsis thaliana* ecotypes, recent duplicates show increased frequency of teM. TeM genes also showed evidence of relaxed selection compared to gbM and unM genes and more frequently associated with the presence-absence variation. Whole-genome duplicates were more broadly expressed compared to SGDs. Duplicate pairs largely retained similar genic methylation patterns and tissue-specificity. We propose that observed differences in genic DNA methylation states reflect biased retention and loss of duplicate genes and may have a role in maintaining gene-dosage balance. Non-CG methylation, in particular, shows strong associations with sequence evolution and may facilitate gene divergence.

Keywords: Gene Duplication, Whole Genome Duplication, DNA methylation, Angiosperms

Introduction

Gene and genome duplication increases organismal gene content and generates a repertoire for functional novelty (1, 2). Whole-genome duplication (WGD), or polyploidy, results in an increase in the entire genomic content of an organism (3) and is more pervasive in plants than other eukaryotic lineages (4–7). At least one WGD event lies at the base of all angiosperms (8) and

subsequent WGDs have occurred at varying frequencies across lineages, except in *Amborella trichocarpa* (9, 10). Small-scale and single-gene duplications (SGDs) also significantly contribute to this gene repertoire (11–14). SGD is a continuous process, with ongoing gene birth and death (15, 16). SGDs can result from a variety of mechanisms, can place genes in new regulatory contexts and chromatin environments, and have been linked to environmental adaptation and domestication in plants (17–21). Following gene duplication, paralogs are retained or lost, a process that is biased depending on the type of duplication and gene function (17, 22–24). Retained paralogs may further diverge in function or expression leading to sub- or neo-functionalization (5). Many, if not most, paralogs, however, will become pseudogenes and are eventually lost (15). Multiple evolutionary and molecular mechanisms contribute to the fate of paralogs (25) and it has been proposed that DNA cytosine methylation may have a role in the evolution of paralogs due to its associations with gene expression (26).

DNA methylation is an important chromatin modification with diverse biological roles (27). Methylation of cytosines in the CG dinucleotide sequence context can be found throughout plants, animals, and fungi; while non-CG methylation occurring at trinucleotide CHG and CHH (where H is A, T or C) contexts is limited to plants (28–31). Non-CG methylation is predominantly found in heterochromatin, alongside CG methylation, and is involved in TE silencing (32). DNA methylation of regulatory regions can affect gene expression (33), however these have not been mapped in many plant species and methylation of upstream regions does not always associate with transcriptional repression (34). Plant genes do show several distinct patterns of DNA methylation within coding regions (henceforth genic methylation) and these have strong associations with gene expression patterns (35). Genes characterized by enrichment of methylated CG sites in the transcribed region and depletion at the transcription start site and transcription termination site are referred to as gene-body methylated (gbM) (33, 36). GbM genes are typically broadly expressed, often conserved between orthologs, and evolve more slowly (37–40). Some genes are methylated similarly to TEs, having both CG and non-CG methylation within coding regions. This transposon-like methylation (teM) is rarely conserved between species and is associated with transcriptional silencing (40–42). The majority of genes in a species are typically unmethylated (unM) in coding regions (40). UnM genes tend to be shorter in length, have fewer exons, and show more variable expression levels compared to gbM genes (38).

Rodin and Riggs (2003) proposed a model of 'epigenetic complementation' in which DNA methylation selectively silences paralogs, leading to expression of these genes in different tissues. As each paralog would be essential to the tissue in which it is expressed and thus complement silencing of the other paralog in that tissue, this could result in retention of duplicate copies and facilitate functional divergence. A similar model was proposed by Adams et al. (43). Alternatively, silencing by DNA methylation may contribute to pseudogenization and loss of paralogs (42, 44). Analyses in humans, mice, and zebrafish found that DNA methylation of duplicate genes corresponds to divergence in gene expression and that more recent duplicate genes were more heavily methylated than older duplicates (45, 46). In plants, DNA methylation was associated with the divergence in duplicated F-box genes in Arabidopsis (44). GbM genes are associated with high expression of paralogs resulting from WGD in soybean, common bean, and cassava (47, 48). Divergence in genic methylation was associated with expression divergence in duplicate genes in potato and tomato (49), while differences in the level of CG methylation between paralogs was suggested to be linked to expression divergence in rice and Arabidopsis (47, 50–53). WGD pairs show lower divergence of CG methylation than SGDs (51). In soybean WGDs

DNA methylation levels and expression diverged with increasing age (50) and that translocation of paralogs to pericentromeric regions and association with TEs resulted in the acquisition of teM, silencing and pseudoization (42).

This complex interaction between the epigenome and duplicate gene evolution remains poorly understood, especially at a broad phylogenetic scale (25, 54). Lineage-specific differences in DNA methylation pathways, rates and histories of gene duplication, and how the data are analyzed can all contribute to differences in the observed associations among DNA methylation and gene duplication. Here, we examine DNA methylation and gene duplication in 43 species from across the angiosperm phylogeny. This analysis shows that there are general trends in the relationship between genic methylation and different types of gene duplication. These patterns correspond to differences in the age of gene duplication, sequence evolution, and breadth of gene expression. It also allows us to identify species and lineages that are outliers to these trends, which may be key to a deeper understanding of these processes.

Results:

Genic methylation patterns across orthogroups

Gene content in angiosperms varies from 'core angiosperm' genes, found in species across the phylogeny, to genes of increasing lineage specificity (9, 55). Retention and loss of paralogs varies within these groups. For instance, it has been observed that certain core angiosperm genes quickly return to single-copy status following WGD (56). As these orthogroups are enriched for certain functions, it is thought that these genes may be dosage-sensitive leading to selection against duplicate copies (24, 55). To understand how genic methylation associates with gene content across angiosperms, we first identified orthogroups from 58 angiosperm species (Table S1). As previously observed (55), orthogroups are bimodally distributed (Figure S1), with the majority represented in a small number of species and a large subset represented across the phylogeny. Following Li et al., orthogroups found in ≥ 85% (51/58) of species were considered 'core angiosperm' (Figure S1). These were further classified as 'core-single copy', if represented by a single gene in ≥ 70% species and the remainder as 'core-other'. Remaining orthogroups were classified based on increasing lineage-specificity: 'cross family' if present in more than one family, 'family-specific' if found in multiple species within a family, or 'species/lineage specific' if limited to a single species.

Genes from 43 angiosperm species (Table S1) were then classified based on their genic methylation patterns (38, 40) as gbM, unM, or teM (Table S2) and examined in the context of our orthogroup classification (Figure 1, Table S3). Across angiosperms, gbM genes are predominantly found in core angiosperm orthogroups, with a median of 37% and 28% in core-single-copy and core-others, respectively. Representation of gbM decreases as othrogroups increase in lineage specificity (Table S3). The exceptions to this are *E. salsugineum*, *B. rapa*, and *B. oleracea;* which are known to be depleted of gbM (57). A large proportion of unM genes are also from core angiosperm orthogroups, but most are from core-other (median 45%) and a lower proportion are single-copy (median 25%) compared to gbM. UnM is increasingly represented in cross-family, family-specific, and species/lineage-specific genes (median 45%, 54% and 44% respectively). Core-angiosperm orthogroups are depleted of teM genes (median 2.5%, and 4.4% in core-single-copy and core-others respectively), while the majority of teM genes are from non-core angiosperm

orthogroups with a higher proportion of lineage/species specific orthogroups (median 24%). This would suggest that teM is biased towards less conserved genes of more recent evolutionary origin. Plotting the percentage of genic methylation across each orthogroup type shows that more than half of the core-single copy and core-other orthogroups were either gbM or unM genes (Figure S2), with the proportion of gbM increasing in core-single copy. From cross-family to lineage-specific orthogroups, the proportion of gbM greatly decreases, teM increases, while unM remains represented throughout.

Gene duplication and genic methylation

The evolutionary fate of whole genome duplicate (WGD) and single gene duplicate (SGD) genes differs (25). Furthermore, SGDs encompass a range of different types of duplication. We identified and classified paralogs as WGDs or one of four types of SGDs (10, 58) (Table S4). Tandem duplicates (TD) are thought to occur through unequal crossing-over, creating clusters of two or more genes adjacent to each other on the same chromosome (14). Proximal duplicates (PD) are gene copies separated by several intervening genes and arose either through the action of local transposition or interruption of ancient tandem duplication (59, 60). Translocated duplicates (TRD; referred to as transposed in some papers) are pairs in which one of the genes is syntenic and the other is non-syntenic (10, 58). The syntenic gene is assumed to be the parent and the other gene the daughter. These duplications can arise either by retrotransposition or DNA-based duplication (61). Finally, dispersed duplicates (DSD) are gene pairs that are neither adjacent to each other, nor collinear, and are produced through various mechanisms (10, 17, 62).

Each class of gene duplication was tested for enrichment or depletion of gbM, unM, or teM (Figure 2, Table S5). WGDs are enriched for gbM in 28/43 and depleted in 6/43 species; enriched for unM in 33/43 and depleted in 4/43 species; and enriched for teM in 3/43 and depleted in 38/43 species. Species with no-significant enrichment/depletion typically still showed the dominant trend of enriched gbM and unM and depleted teM. The three Brassicaceae species depleted in gbM showed no enrichment of gbM in WGDs. *Solanum tuberosum* is the only species that shows a complete opposite trend to all other species, with WGDs being depleted in gbM and unM and enriched in teM. Commercially cultivated *S. tuberosum* is a very recent autotetraploid (63) and the methylome data are from tetraploid plants (49), which may be a factor in its exceptionalism.

SGDs show different patterns between types of SGD, however, there seems to be a more general split between local (tandem and proximal) and distal (translocated and dispersed) duplicates. Tandem and proximal genes were depleted of gbM genes in all species except the three Brassicaceae (40/43 species). Tandem genes are enriched for unM in 40/43 and depleted in 3/43 species, while proximal genes are enriched in 33/43 and depleted in 5/43 species. TeM methylation varies. Tandem genes are enriched for TeM in 18/43 and depleted in 18/43 species. Proximal genes show more enrichment than tandem for teM, being enriched in 29/43 and depleted in 3/43 species. Distal duplicates are generally enriched for teM and depleted in gbM and unM. Dispersed genes were enriched for teM in all species except *S. tuberosum*, which showed depletion. Translocated genes were enriched for teM in 37/43 and depleted in 3/43 species. Dispersed genes showed no enrichment for unM and were depleted in 38/43 species, while translocated genes were enriched in 9/43 and depleted 20/43 species. For gbM, dispersed genes were enriched in 9/43 and depleted in 23/43 species, while translocated genes showed no enrichment of gbM and depletion in 30/43 species. Enrichment of teM amongst distal duplicates and the greater number of species with

enrichment of teM in proximal vs tandem duplicates suggests that teM becomes more common as genes move to differing sequence or chromatin environments.

We tested these patterns of enrichment and depletion to see if there were any effects of phylogeny on the observed pattern. We used Pagel's lambda (λ) (64, 65), which ranges from 0 (no phylogenetic signal) to 1 (strong phylogenetic signal). Phylogenetic signals were not observed in any of the patterns except WGD gbM genes (λ ~ 0.78 ; Table S5). This result was still significant even after the removal of the three Brassicaceae species (λ ~ 0.79).

Similarity and divergence of genic methylation between paralogs

Models of how DNA methylation affects duplicate gene evolution presume that paralogs differ in DNA methylation. We assessed how frequently paralogs differ in genic methylation for each type of duplication (Figure S3, Table S6, S7). WGD pairs tend to resemble each other, with ~70% to 97% gene pairs in a species (median - 84%) having the same genetic methylation profiles. SGD pairs also predominantly shared the same genic methylation profile: tandem (~69% - 93%, median – 82%), proximal (~65% - 90%, median – 78%), and dispersed (~65% - 91%, median – 77%). Translocated duplicates had the broadest range (~51% - 90%, median – 75%), but for all species the majority of gene pairs showed similar genic methylation profiles (Table S6, S7). So regardless of the type of duplication, most duplicate copies retain similar DNA methylation profiles, however, divergence is not rare and seems slightly more frequent in certain SGDs.

We typically cannot ascertain the original genic methylation state or the directionality in which changes occur, however, for translocated pairs it is possible to identify a parental and translocated copy, as the parental gene has maintained synteny (58). A higher proportion of translocated copies were teM genes compared to their parental gene in all species except *C. sativus*, *P. halli* and *T. cacao* (Table S8). The median proportion of teM genes increased from 14% in parental copies to 19% in the daughter copies of the translocated pair while the proportion of gbM genes decreased from 27% to 20% (Table S9). Translocated copies also showed a lower proportion of gbM (39/43 species) and unM genes (28/43 species) compared to their parental copies in most species (Table S8, S9).

Genic methylation and duplicate age

Differences in retention or loss of paralogs can result in biases in the age of duplicate pairs and it has been proposed that silencing or 'expression reduction' by DNA methylation can facilitate paralog retention (46). Synonymous substitution (Ks) distributions have been widely used to date gene duplication events, as synonymous substitutions are often assumed to evolve neutrally and therefore can accumulate with time (15, 16). We examined the Ks distributions for WGDs and SGDs based on the genomic methylation of each duplicate pair (Figure 3; Figure S4; Table S11, S12). We did not split SGDs based on type of duplication, as similar patterns were observed for all four types of SGDs. To observe larger trends across the phylogeny, we ordered groups of duplicate pairs based on genic methylation and median Ks (Figure 3A,B, Table S11). WGD pairs in which one or both of the genes are gbM often have lower Ks values and this is most evident in gbM-gbM pairs. UnM genes, especially those in unM-teM or unM-unM pairs usually have higher Ks. This would suggest that retained unM-containing WGD pairs are typically older while gbM-containing WGD pairs are typically younger. No obvious trend is observed for teM-containing

WGDs, except that there is a cluster of species with low Ks teM-teM pairs (Figure 3A). However, this may be a spurious result due to the depletion of teM in WGD and scarcity of teM-teM pairs. A notable exception to the trends in WGD is *B. vulgaris*, where gbM-gbM pairs have the highest Ks and unM containing pairs have lower Ks. The last known WGD event in *B. vulgaris* was the core-eudicot γ WGT (10, 66, 67) and *B. vulgaris* has the fewest identified WGDs in our analysis (587) other than *A. trichopoda* (14) (Table S4).

For SGDs, teM-containing duplicate pairs typically have much lower Ks values. This is most evident amongst teM-teM pairs, but gbM-teM and unM-teM typically show this same trend. No obvious trend exists for gbM or unM genes in SGD pairs, although these are usually shifted toward higher Ks values than teM genes (Figure 3B). This suggests that teM SGD genes are typically more recent in origin. However, there are notable exceptions. Both *M. truncatula* and *P. x. Bretschneideri* have the exact opposite trend, with teM SGD pairs having higher Ks values. We further tested this relationship between SGD age and genic methylation for translocated genes using a method independent of the Ks-based approach. As the syntenic gene is assumed to be parental for translocated copies, the daughter gene can be parsed into different periods (epochs) since speciation, by sequential exclusion to the closest outgroup, as employed in *MCScanX-transposed* (58) (Table S13). We observe that more recent translocated duplicates, particularly those specific to that species, were enriched in teM genes and depleted in gbM and unM genes, while more ancient translocated duplicates were depleted of teM genes and enriched for gbM and unM genes (Figure S5, Table S14), confirming our results from the Ks analysis.

Genic methylation and sequence evolution

We then compared the ratio of non-synonymous (Ka) to synonymous (Ks) substitutions (Ka/Ks) to observe differences in the evolution of paralogs based on their genic methylation (Figure 4, Figure S6, Table S15, S16). As done for Ks, we ordered duplicate pairs based on their genic methylation and median Ka/Ks to observe general trends across species (Figure 4A,B). The vast majority of duplicate pairs have a Ka/Ks < 1, which is indicative of purifying selection. This was generally true regardless of genic methylation. However, for both WGD and SGD genes, teM containing duplicate pairs, in particular teM-teM pairs, have higher Ka/Ks values and gbM, especially gbM-gbM pairs, have lower Ka/Ks. This suggests that teM genes, especially teM-teM pairs are under relaxed selective constraints. Interestingly, the distribution of teM genes in many species also shows a shift toward Ka/Ks > 1, which may suggest positive selection, although much more rigorous testing is necessary (Figure 4, Figure S6).

Ongoing gene duplication and loss within a population will create copy number and presence absence variation (PAV) within a species. We examined the relationship between genic methylation and PAVs in four species (*B. oleracea*, *S. lycopersicum*, *Solanum tuberosum*, and *Z. mays*) with available PAV data (Figure S7; Table S17). When examining all genes (duplicated or not), teM genes showed a significant enrichment amongst PAVs (FDR adjusted $p < 0.001$) in all four species, while gbM showed significant depletion except for *B. oleracea*, where it was not statistically significant. UnM genes showed conflicting patterns, being enriched in *S. lycopersicum* (FDR adjusted $p < 0.001$), and depleted in the other three species. These same observations held up when the analysis was limited to duplicate genes (Figure S7, Table S17).

Genic methylation and gene expression divergence

Differentiation of gene expression between duplicate pairs is common (10, 17, 68) and would be the primary means by which DNA methylation affects paralog evolution. We explored how genic methylation relates to expression differentiation using gene expression atlases in *A. thaliana*, *G. max*, *P. vulgaris* and *S. bicolor* (69–71). We first correlated expression levels of paralogs across conditions, observing that paralogs have different distributions of correlations based on their genic methylation (Figure S8). GbM-gbM pairs are typically positively correlated. Other duplicate pairs show a range of distributions, most being bimodal with a prominent peak around 0 and another closer to 1, however there are notable exceptions. GbM-unM pairs have very different distributions of correlations based on species, from predominantly positively correlated in *G. max* and *P. vulgaris* to largely uncorrelated in *S. bicolor*. *G. max* paralogs are more positively correlated than other species, regardless of methylation status. *G. max* has the most recent history of WGD and the most WGDs of the four species, having had a polyploid event ~13 MYA that is not shared by its relative *P. vulgaris*, which may explain the higher positive correlations between paralogs.

To further examine these relationships, we calculated the specificity of expression for each gene using τ (Tau) (72). The value of τ ranges from '0' (broad expression) to '1' (narrow expression). Collectively, gbM genes have the lowest τ, teM the highest, while unM genes have a breadth of intermediate values between gbM and teM (Figure 5A, S9). Examined by duplication type, WGDs have lower τ, local SGDs (tandem and proximal) have the highest τ, while translocated and dispersed are usually intermediate between WGD and local SGDs (Figure S10). However this pattern amongst duplication types varies when broken down by genic methylation. UnM paralogs typically resemble the overarching trend. This is not surprising as most genes are unM. GbM duplicates all have lower τ, and there is little difference between WGD and distal SGD, while local SGDs show different distributions depending on the species. TeM genes all have higher τ, although WGD teM genes have lower values compared to SGDs. Duplicate pairs with the same methylation profiles typically have lower absolute difference in τ than duplicate pairs that differed in genic methylation (Figure 5B, S11) and the greatest differences in τ were observed for gbM-teM pairs, followed by gbM-unM.

To further understand these differences in expression specificity, we examined the distribution of τ for gbM, teM, and unM genes separately, subsetting these based on the methylation of their duplicate pair (Figure 5C-E, S12). So in the case of gbM genes, we compared the distribution of τ for the gbM gene in gbM-gbM, gbM-teM, and gbM-unM pairs. In both *A. thaliana* and *S. bicolor*, the gbM gene in gbM-teM and gbM-unM genes had a higher τ than those in gbM-gbM pairs. In other words, in these species, a gbM gene that has a teM or unM duplicate pair, often has a narrower range of expression compared to other gbM genes. We did not observe this same trend in either *G. max* or *P. vulgaris* gbM genes, although *G. max* gbM-unM had slightly higher τ (Figure S12). TeM genes, in which the other duplicate was gbM, had lower τ than teM genes from teM-teM pairs or teM-unM pairs in all species. We note that despite the greater difference between gbM-teM pairs compared to gbM-gbM pairs (Figure 5C-E, S12), many gbM-teM pairs do not show large differences in τ.

Distributions of genic methylation and transposons association

DNA methylation in genes is shaped by a number of factors. DNA methylation tends to be higher in the centromere and pericentromeric regions (73, 74) and non-CG methylation of genes is associated with local TE content and genome size (39, 40). In *G. max*, teM genes were enriched in pericentromeric regions (42) and it was suggested that translocation of duplicates to pericentromeric regions was a major determinant of teM. To examine how genic methylation is distributed throughout plant genomes, we used the number of genes, number of TEs, and number

of nucleotides from TEs (TE-base pairs) in sliding windows as a proxy for regions of euchromatin and heterochromatin and correlated these with the number of gbM, unM, and teM genes in those windows (Figure S13, S14, S15, Table S18). GbM, unM, and teM all showed positive correlations with the distribution of genes, except for *A. thaliana* teM genes, which were negatively correlated (Pearson's r = -0.30, FDR adjusted $p < .001$). The distribution of gbM and unM genes was negatively correlated with TE content across the majority of species (gbM: 34/43 species; unM: 30/43 species). Species with a positive correlation between gbM/unM genes and TE content also have a positive correlation between TE content and gene distribution (Figure S13). TeM genes were positively correlated with TE distribution in most species (28/43), but surprisingly, a number of species did not show this expected pattern. These trends were similar when the analysis was limited to duplicate genes (Table S18).

Many of the species deviating from the expected pattern of teM distribution were in the Fabaceae (legumes) and Poaceae (grasses), suggesting phylogenetic effects (Table S19). Phylogenetic signals were found for the correlation between teM genes and TEs ($\lambda = 0.73$, FDR adjusted $p = .004$) and between gbM genes and TEs ($\lambda = 0.82$, FDR adjusted $p = .006$). We retested the gbM-TE correlation with *E. salsugineum*, *B. rapa*, and *B. oleracea* removed due to the loss of gbM in these species (57), and still found significant evidence of phylogenetic signal for gbM genes and TEs ($\lambda = 0.79$, FDR adjusted $p = .01$). As genome size correlates with genic non-CG methylation (38, 40), we also tested for phylogenetic signals using genome size, but found none, suggesting that the lineage-specific changes in distribution of gbM and teM genes is unlikely to be derived from genome size variation

The global distribution of genes ignores local differences in TE content. To address this, we calculated the presence of TEs in the coding region and 3 kb upstream and downstream of paralogs. Compared to the genome average, teM genes showed enrichment for genes overlapping TEs in 30/43 species, while gbM and unM genes were mostly depleted (19/43 and 22/43 species respectively). Interestingly, some members of the legume and grass family did not show significant enrichment for TE presence in teM genes (Figure S16, Table S20), further confirming the chromosomal level lineage-specific changes observed in TE and gene distribution in these two families (Figure S13). Similarly, WGDs showed depletion for genes with overlapping TEs in 31/43 species, while SGDs mostly showed enrichment (Figure S14, Table S21).

Population scale genic methylation of duplicate genes

Within a species DNA methylation can vary across the population (75–79) which could lead to differences in paralog evolution across the population. We examined this using methylomes from 928 *A. thaliana* accessions in the 1001 Epigenomes Project (76). Genes were binned based on the frequency of genic methylation classification in the population (0%, <25%, 25%-50%, 50%-75%, >75%). To obtain high frequency across the population, it seems more parsimonious that genic methylation patterns were established early and are stable enough to spread with *A. thaliana*, rather than to be established independently multiple times. Low frequency genic methylation, could be derived, resulting from either a recent gain/loss of DNA methylation within a subpopulation, e.g. a recent TE insertion resulting in gain of teM. Alternatively, that state could be ancestral, with derived genic methylation states occurring relatively early and then spreading to higher frequency. No trend in Ks was observed for either gbM or unM paralogs. Ka/Ks is higher in low frequency than high frequency gbM paralogs, while for unM paralogs it increases slightly with higher unM

frequency. For teM, increasing frequency within the population corresponds to a decrease in Ks and an increase in Ka/Ks (Figure 6). This would suggest that teM was established early enough in younger paralogs to spread to high frequency.

Discussion

Genome evolution is shaped by a confluence of multiple factors. We have examined the relationship of two of these factors, gene duplication and DNA methylation. We find that WGDs are generally enriched in gbM and unM, while being depleted in teM. SGDs are typically depleted in gbM and show differential enrichment of unM and teM based on the type of duplication. Local SGDs (tandem and proximal) are enriched for unM and variably enriched/depleted for teM. Distal SGDs (translocated and dispersed) are often depleted of unM and enriched for teM. While notable exceptions to these trends exist, we suspect these exceptions hint at more interesting underlying biology. For example, the notable depletion of gbM in WGDs of certain Brassiceae can be explained by the loss of gbM in these species (57). A similar trend is observed in the Cucurbitaceae, and while no similar loss of gbM is known in these species, this and other exceptions to these trends warrant further investigation.

Enrichment or depletion between types of duplication could occur either through shifts in DNA methylation states or biased retention and loss of one group of genes. WGD would result in the duplication of all genes, including teM, and would not be expected to result in changes to the local sequence and chromatin environment of paralogs. We find that despite differing histories of WGD, the majority of WGD pairs in all species shared the same genic methylation status. In line with this, studies of synthetic allopolyploids in both *Mimulus* and *Brassica* do not show extensive changes to genic methylation patterns, which would suggest that the parental state is largely maintained following allopolyploidy (80, 81). In the *Brassica* synthetic allopolyploids generational changes were observed to CHH methylation in flanking regions, however, it is unclear if these corresponded to expression differences (81). In contrast to WGD, SGD is a continual process, capable of selectively amplifying specific gene families. It also places genes in potentially new sequence and chromatin environments. Extensive switching of genic methylation was previously observed for translocated duplicates in *G. max*, where it was proposed that movement to the pericentromere resulted in silencing (42). We do find some evidence for this. Translocated duplicates had the greatest percentage of pairs differing in genic methylation and teM genes were enriched in TE-rich regions. However, the majority of SGD pairs, including translocated duplicates, in all species have the same genic methylation status. We also observe lineage-specific differences in the distribution of teM genes that would suggest that this process is much more complex. In some instances, changes in genic methylation could be masked through convergence in methylation states by *trans*-acting mechanisms, such as RNA-directed DNA methylation (82). This was observed in the case of the *A. thaliana PAI* gene family (83). However, the simplest explanation is that paralogs retain the same genic methylation state as the parental gene in most cases.

Enrichment and depletion of genic methylation between orthogroups and types of duplication is therefore most likely explained by association of DNA methylation with different classes of genes and subsequent biased retention or loss following duplication. Retained WGDs are biased towards certain functional categories (17). This is explained by the gene balance hypothesis, which postulates that dosage-sensitive genes, especially those of multimeric complexes, will be

preferentially retained following WGD to retain dosage balance between genes (84). At the same time, it has been observed that a subset of core angiosperms orthogroups exhibit a different type of dosage-sensitivity and return relatively quickly to single-copy status (24, 56). SGDs show different biases and are often associated with functions relevant to environmental responses and adaptation (85). Our observations of gbM duplicate gene sequence evolution and expression, as well as broader observations of gbM conservation and function (37–40), fits with the expectations of the gene-balance hypothesis and observations of single-copy core angiosperm orthogroups. These same characteristics would also explain their depletion in SGDs. If gbM genes are dosage-sensitive, then duplicates of an SGD would be selected against and eliminated. UnM paralogs are harder to place, seeming intermediate between gbM and teM in most aspects. They are the largest of the three groups and broadly represented across core angiosperm orthogroups and more lineage-specific orthogroups. Many transcription factors and kinases have more tissue-specific expression, characteristic of unM, and are retained following WGD (86). In species that have lost gbM, the gbM ortholog is unM (57). UnM might be considered the 'default' state and capture a range of classes from gbM-like to teM-like genes. At the same time, tandem and proximal duplication would be more likely to place a duplicate in the same chromatin context as its parent and are also often associated with environmental adaptation (17).. These factors would favor the retention of unM in both WGD and local SGDs.

The narrow expression, higher Ka/Ks ratios, and enrichment in PAV of teM paralogs suggest that these are on the path to pseudoization and the dustbins of evolution. This would lead to their general depletion in WGDs. While most SGD teM paralogs likely face the same fate, continual generation of new SGDs and the more frequent association of SGDs with TEs will provide a constant source of new teM paralogs. This would explain their enrichment and also explain the younger evolutionary age of teM pairs. Increasing frequency of teM within the *A. thaliana* population is associated with a lower Ks, showing that teM is more frequent in recent duplicates and rare amongst older duplicates. To achieve wide-spread frequency in the population, the simplest explanation is that in these cases, teM was established early following duplication and subsequently spread with *A. thaliana*. For older duplicates, the lower frequency could arise either from subsequent loss of teM and conversion to gbM or unM or from infrequent conversion of gbM and unM to teM. We suspect that the latter is more common. It has been recently observed that a subset of gbM genes can shift to teM due to DNA methylation homeostasis (87). We find that for gbM-teM duplicate pairs, the gbM paralog is more narrowly expressed in both *A. thaliana* and *S. bicolor* than compared to gbM-gbM pairs. Corresponding to this, we find that the teM paralog in gbM-teM pairs is more broadly expressed than compared to teM-teM pairs. Hints of this are also observed for unM-teM pairs. This may suggest that duplicates that shift to teM are more predisposed to do so to begin with. It is possible that some teM genes are misannotated transposons (88, 89), however, teM has been found in many known protein-coding gene families (44), including species-specific genes (90). Furthermore, for translocated duplicates, the parental copy is a syntenic gene (58), which would further support it as an actual duplicated gene.

It has been proposed that silencing by DNA methylation can result in retention of paralogs and their functional divergence (e.g. epigenetic complementation) (26, 43, 46). Alternatively, it is argued that silencing leads to pseudogenization and gene loss (42, 44). Neither hypothesis is necessarily wrong or exclusive to the other. Our results suggest that pseudogenization and loss is the predominant consequence. However, there is also suggestive evidence for epigenetic complementation. Many teM containing duplicates have a Ka/Ks > 1, which might hint at positive

selection. Rapid functional divergence of SGDs was observed in grasses and many of these have characteristics similar to what we observe in teM SGDs (91). While the original epigenetic complementation model envisioned DNA methylation as the mechanism, H3K27me3 or other chromatin marks may be more relevant. These have been understudied in the context of gene duplication. DNA methylation in *cis*-regulatory regions (CREs) can also have an effect (92). However genome-wide maps of CREs remain incomplete for most plant species. Our results establish broad genome-wide trends in the relationship between genic methylation and gene duplication, however, the specific effects of each gene will require individual examination.

Methods:

DNA methylation analyses

Published whole-genome bisulfite sequencing (WGBS) from 43 Angiosperm species (See Supporting Information, Table S1) were mapped to their respective genomes using methylpy v1.2.9 (93). For all analyses, only the primary transcript was used. A background methylation rate was calculated for CG, CHG, CHH, and non-CG (combined CHG & CHH) methylation by averaging the percentage of methylated sites in each context across coding regions of all species (40). Each gene was then tested for enrichment of CG, CHG, CHH, non-CG in its coding region against this background rate using a binomial test. P-values were corrected for false discovery rate (FDR) by the Benjamini-Hochberg (BH) procedure (94). Genes enriched for CG methylation with ≥ 10 CG sites and non-significant CHG or CHH methylation were classified as gene-body methylated (gbM). Genes enriched for CHG, CHH, or non-CG and ≥ 10 sites in that context were classified as TE-like methylated (teM). Genes with ≤ 1 methylated site in any context or a weighted methylation (95) for all contexts (CG, CHG, or CHH) ≤ 2% were classified as unmethylated (unM). Genes lacking DNA methylation data were considered 'missing' and those with intermediate DNA methylation levels not fitting the above criteria as 'unclassified'.

Orthogroups analyses

Protein sequences from 58 angiosperm species (See Supporting Information, Table S1) were classified into orthogroups using Orthofinder v2.5.2 (96, 97), with the options '-M dendroblast -S diamond_ultra_sens, -I 1.3'. Accounting for missing annotations, orthogroups represented in ≥51 species (~87.9%, Figure S1) were classified as "core angiosperm" orthogroups. This is equivalent to Li et al. who used 32/37 species (~86.5%) (55). Following Li et al., we further classified core angiosperm orthogroups as single-copy if they were represented by a single gene in 70% or more species. Non-core orthogroups were classified as "cross-family" (present in ≥ 2 species belonging to different families), family-specific (present in ≥ 2 species within the same family), and "lineage/species-specific" (present in only one species).

Gene duplication classification

For each species, DIAMOND (98) was used to perform a blastp of all genes against itself and *A. trichopoda*. Hits with an e-value < 1e-5 were retained. Many hits belonged to different orthogroups which could confound downstream analyses, so these were further filtered to remove genes from different orthogroups. Duplicate genes were classified by *DupGen_finder-unique (10)*, requiring ≥ 5 genes for collinearity and ≤ 10 intervening genes to classify as 'proximal' duplicates.

MCScanX-transposed (58) was used to detect translocated duplicates occurring within different epochs since species divergence (Table S13). Enrichment between type of duplication and genic methylation was determined by a two-sided Fisher's exact test (99) with FDR-correction by BH and plotted using *heatmap.2* in the R package *gplots* (100). The phylogenetic tree in Figure 1 was created using the R packages 'V.PhyloMaker' (101) and 'phytools' (102).

Sequence evolution

The calculate_Ka_Ks_pipeline.pl (10) was used to determine nonsynonymous (Ka) and synonymous substitutions (Ks) for duplicate pairs. Protein sequences are aligned by MAFFT (v7.402) (103) and converted into a codon alignment with PAL2NAL (104). KaKs_Calculator 2.0 is then used to calculate Ka, Ks, Ka/Ks using the γ-MYN method (10, 105). We tested the distribution of Ks and Ka/Ks for duplicate gene pairs for divergence from the distribution of an equal number of randomly selected genes using the Kolmogorov-Smirnov test (106) with FDR-correction by BH. PAV variants were downloaded for *B. oleracea (107)*, *S. lycopersicum (108)*, *S. tuberosum (109)*, and *Z. mays (110)*. Only genes present in the reference genome were considered. For *S. tuberosum* and *Z. mays,* genes with an average read coverage of < 0.2 in ≥ 1 accession were considered PAV. Enrichment was tested for using a two-sided Fisher's Exact test with FDR-correction by BH.

Gene expression

Expression data for *A. thaliana*, *G. max*, *P. vulgaris*, and *S.bicolor* were taken from published gene expression atlases (69–71, 111). Raw data for *A. thaliana* was downloaded from NCBI SRA (PRJNA314076 and PRJNA324514), mapped using STAR (112), and gene counts normalized by DESeq2 (113). For *G. max*, *P. vulgaris*, and *S.bicolor*, normalized counts were downloaded from Phytozome (114). Pearson correlation coefficients were calculated for each duplicate pair and the tissue-specificity index (τ) calculated for each gene, as described in Yanai et al. (72).

Transposons and genomic distribution

TEs were called *de novo* for all species using the Extensive *de-novo* TE Annotator pipeline (115). We calculated the total number of genes, genes belonging to each of the genic methylation classes, the number of TEs, and number of TE nucleotides in 100 kb sliding windows with 50 kb steps. Pearson correlation coefficients were calculated using the '*rcorr*' function in '*corrplot*' (116). Genome sizes were obtained from the Plant DNA C-value database (117) (release 7.1).

Phylogenetic signal

Enrichment/depletion of genic methylation, correlations and genome sizes were tested for phylogenetic signal using Pagel's lambda with the function *phylosig* in the R package *phytools (102)*. Pagel's lambda measures phylogenetic signal under a Brownian motion model of trait evolution (64, 65). The input phylogenetic tree and branch lengths (Dataset S1) was generated with orthofinder, as described above (97). For enrichment/depletion of genic methylation in different duplication types, statistically significant depletion was coded as -1, enrichment 1, and non-significant results 0.

Arabidopsis diversity

Aligned WGBS data from the *Arabidopsis thaliana* 1001 Epigenomes Project (76) was downloaded from the Gene Expression Omnibus (GEO Accession GSE43857). 928 accessions were previously aligned by methylpy and compatible with our methods. Genes were classified as before using and the frequency of each genic methylation class for each gene in the population calculated. Ks & Ka/Ks data were the same as those computed above for duplicate pairs.

Data availability and research reproducibility

Data used in this study are listed in the Supporting Information and Table-S21. Formatted genomes and annotations suitable for reproducing our results are available at DataDryad (TO BE ADDED AT PUBLICATION) . Conda environments and code to reproduce these results are available at: https://github.com/niederhuth/DNA-methylation-signatures-of-duplicate-gene-evolution-in-angiosperms.

Acknowledgements: We would like to thank Dr. Patrick Edger for use of the unpublished *C. violaceae* genome and Dr. Leslie Kollar for reviewing the manuscript. The authors would like to thank and dedicate this work to the healthcare workers during the COVID-19 pandemic. Dr. Niederhuth also dedicates this work to his son Harrison Edgar Niederhuth who was born during the writing of this manuscript. This work was supported by Michigan State University, by USDA National Institute of Food and Agriculture Hatch Funds (project number MICL02572), and the National Science Foundation (grant IOS-2029959). S Marshall Ledford was supported by the Plant Genomics@MSU REU (National Science Foundation grant DBI-1757043).

Author contributions: S.K.K.R and C.E.N conceptualized and designed the work and analysis. S.K.K.R, C.E.N. and S.M.L performed data analysis. S.K.K.R made figures and wrote the original draft; C.E.N wrote and edited the manuscript. All authors read and approved the final manuscript.

Figure Legends:

Figure 1: Distribution of orthogroups in genic methylation classes. For each species, the percentage of genes classified into different orthogroup categories (core-single copy, core-other, cross-family, family-specific, and species/lineage-specific) in each of the three genic methylation classification (gbM, teM, and unM genes).

Figure 2: Patterns of genic methylation across different types of gene duplicates. Enrichment or depletion of each genic methylation class (gbM, teM, and unM) for each type of gene duplication (WGD, tandem, proximal, translocated, and dispersed). Increasing shades of cyan indicates greater depletion, while increasing shades of magenta represents greater enrichment. Unless indicated, all associations are statistically significant at a FDR-corrected p-value < 0.05. 'NS' indicates no statistical significance.

Figure 3: Relationship between genic methylation and the age of gene duplication. Bar plots showing the number of species in each of the duplicate-pair genic methylation classifications (gbM-gbM, gbM-teM, teM-teM, unM-unM, gbM-unM, and unM-teM) ranked based on median Ks values (synonymous substitutions) for whole-genome duplicates (A) and single-gene duplicates

(B). Density plots (C and D) and box plots (E and F) show the distribution of synonymous substitutions (Ks) for each of the duplicate-pair genic methylation classifications in *B. distachyon* and *P. vulgaris* respectively. The black line in the density plots represents the Ks distribution of all duplicate pairs.

Figure 4: Relationship between genic methylation and sequence evolution for duplicate pairs. Bar plots showing the number of species in each of the duplicate-pair genic methylation classifications (gbM-gbM, gbM-teM, teM-teM, unM-unM, gbM-unM, and unM-teM) ranked based on median Ka/Ks values (ratio of Ka, non-synonymous substitutions to Ks, synonymous substitutions) for whole-genome (A) and single-gene duplicates (B). Density plots (C and D) and box plots (E and F) show the distribution of Ka/Ks ratios for each of the duplicate-pair genic methylation classifications in *B. distachyon* and *P. vulgaris* respectively. Dotted line at Ka/Ks ratio of '1' suggestive of neutral selection. Black line in the density plots represents the Ka/Ks distribution of all duplicate pairs.

Figure 5: Gene expression specificity of *A. thaliana* duplicate gene pairs. Tissue-specificity index, Tau (τ), ranges from 0 (broadly expressed) to 1 (narrowly expressed). (A) Tissue specificity of genes based on genic methylation classification (gbM, unM, and teM). (B) Absolute difference in tissue-specificity index (τ) between pairs of duplicate genes with similar or divergent methylation. Differences in Tau specificity of gbM, unM, and teM genes (C, D, and E respectively) when the other duplicate pair has the same or a different genic methylation status. For example, for gbM genes, the tau specificity was plotted for all gbM genes and the gbM paralog in gbM-gbM, gbM-teM, and gbM-unM pairs. For unM genes, the tau of only the unM paralog is shown and similarly for teM genes, only the tau of the teM paralog is shown.

Figure 6: Association of genic methylation frequency with the age of gene duplication and sequence evolution in 928 *A. thaliana* accessions. A) Density plots showing the Ks distribution of genes at different frequencies (0%, <25%, 25%-50%, 50%-75%, >75%) in the population for gbM, unM, and teM genes. Boxplots of Ks (B) and Ka/Ks distributions (C) for gbM, unM, and teM genes at different frequencies.

Supplementary figures

Figure S1: Distribution of orthogroups across 58 angiosperm species. Histogram showing the number of orthogroups represented in 1 to 58 species (A) and the same plot zoomed into species with 2 to 58 species (B). Orange colored bars represent those orthogroups classified as 'core angiosperm'.

Figure S2: Distribution of genes classified as gbM, unM, teM, unclassified, and 'missing methylation data' across different orthogroup classifications (Core-single copy, core-others, cross-family, family-specific, and species/lineage specific).

Figure S3: Proportion of paralogs with similar and divergent DNA methylation profiles. The proportion of duplicate pairs with similar DNA methylation profiles among different types of duplicate genes (Whole-genome duplicates - WGD, Single-gene duplicates - tandem, proximal, translocated, and dispersed) are shown in blue. Yellow bars represent the proportion of duplicate pairs with divergent DNA methylation profiles. Grey bars represent cases where DNA methylation status of at least one of the duplicate pairs was 'undetermined'.

Figure S4: Distribution of genic methylation classified genes based on synonymous substitution (Ks) across different types of gene duplicate pairs. Whole-genome duplicates - WGD, Single-gene duplicates - SGD (combined data from tandem, proximal, translocated, and dispersed duplicates).

Figure S5: The percentage of gene copies in each genic methylation class for translocated genes that have duplicated during that 'epoch' since divergence from the species on the x-axis. For example, in *A. duranensis* translocated genes that have duplicated since *A. duranensis* diverged from *A. ipaensis* are shown on the x-axis under *A. ipaensis*. Those shown under *G. max*, duplicated in the period since the common ancestor of *A. duranensis* and *A. ipaensis* diverged from their common ancestor with *G. max*, but before the divergence of *A. duranensis* and *A. ipaensis*. Horizontal dotted lines indicate the percentage of each genic methylation class in all translocated duplicates. Bars above this line indicate enrichment, below this line depletion.

Figure S6: Distribution of genic methylation classified genes based on the ratio of nonsynonymous substitution (Ka), with synonymous substitutions (Ks) across different types of gene duplicate pairs. Whole-genome duplicates - WGD, Single-gene duplicates - SGD (combined data from tandem, proximal, translocated, and dispersed duplicates).

Figure S7: Percentage of Total (all genes), gbM, teM, and unM genes with known presence-absence variations. This plot was not restricted to duplicate genes, however the same results were found when limited to duplicates (Table S16). A two-sided Fisher's Exact Test was used to test for depletion or enrichment of PAVs amongst each category of genic methylation. *FDR corrected p-value < 0.05, **FDR corrected p-value < 0.01, ***FDR corrected p-value < 0.001, NS – Not significantly different.

Figure S8: Distribution of gene expression correlations of duplicate pairs based on genic methylation (gbM-gbM, gbM-teM, teM-teM, unM-unM, gbM-unM, and unM-teM) in *A. thaliana*, *G. max*, *P. vulgaris*, and *S. bicolor*.

Figure S9: Tau specificity of gbM, teM, and unM genes in *G. max*, *P. vulgaris*, and *S. bicolor*.

Figure S10: Tau specificities of different types of duplicate genes in *A. thaliana*, *G. max*, *P. vulgaris*, and *S. bicolor*. The distribution of tau for gbM, unM, and teM genes is shown for all duplicates and also broken down based on the type of duplicate gene.

Figure S11: Absolute differences in Tau specificity between duplicate pairs in *G. max*, *P. vulgaris*, and *S. bicolor*. Data is broken down based on the genic methylation of the duplicate pairs (gbM-gbM, gbM-teM, teM-teM, unM-unM, gbM-unM, and unM-teM).

Figure S12: Distribution of Tau specificities for gbM, unM, and teM genes separated based on the methylation of their duplicate pair for *G. max*, *P. vulgaris*, and *S. bicolor*. For example, for gbM genes, the tau specificity was plotted for all gbM genes and the gbM paralog in gbM-gbM, gbM-teM, and gbM-unM pairs. For unM genes, the tau of only the unM paralog is shown and similarly for teM genes, only the tau of the teM paralog is shown.

Figure S13: Pearson correlations between genic methylation classes (gbM, unM, teM) and genomic features (number of genes, TEs, and TE nucleotides). Number of genes, TEs and TE nucleotides were calculated in 100kb sliding windows with a 50kb step size. Increasing shades of cyan indicate negative correlation, while increasing shades of magenta represent positive correlations. Unless indicated, all associations are statistically significant at a FDR-corrected p-value < 0.05. 'NS' indicates no statistical significance.

Figure S14: Correlations between number of genes, TEs, and TE nucleotides, and different genic methylation classifications (gbM, unM, and teM) plotted separately for each species. Increasing blue indicates a positive correlation, increasing red indicates a negative correlation. Boxes marked with an 'X' are statistically insignificant (p-value > 0.001).

Figure S15: Distribution of genic methylation classified genes and genomic features across the largest chromosomes in representative species: *A. thaliana*, *G. max*, and *Z. mays*.

Figure S16: Enrichment and depletion of transposable elements (TEs) with gbM, teM, and unM paralogs in each species. TEs within 3 kb upstream, downstream or within the gene body were considered associated with that gene. Fisher Exact test odds ratio of less than 1 represents depletion (represented in shades of cyan), greater than 1 indicates enrichment (represented in shades of magenta). Unless indicated, all associations are statistically significant at a FDR-corrected p-value < 0.05. 'NS' indicates no statistical significance.

Figure S17: Enrichment and depletion of transposable elements (TEs) with different types of duplication in each species. TEs within 3 kb upstream, downstream or within the gene body were considered associated with that gene. Fisher Exact test odds ratio of less than 1 represents depletion (represented in shades of cyan), greater than 1 indicates enrichment (represented in shades of magenta). Unless indicated, all associations are statistically significant at a FDR-corrected p-value < 0.05. 'NS' indicates no statistical significance.

Supplementary tables

Table S1: Genomes, methylomes, and mapping statistics for data used in the study.

Table S2: Classification of genic methylation of all genes in each species. A binomial test was applied to classify genes into gene body methylated (gbM genes), transposable element-like methylated (teM genes), and unmethylated (unM genes). All other genes were classified as either 'Unclassified' or 'Missing' if methylation data was not available for that gene.

Table S3: Enrichment and depletion of genic methylation classified genes across different orthogroup classifications. Fisher exact test odds ratios of enriched associations are colored in green, depleted associations are in brown. FDR adjusted p-value < .05 are indicated in light blue.

Table S4: Number of genes derived from different types of duplications in each species. Genes were classified into each type of duplication using the Dup-Gen_finder-unique pipeline.

Table S5: Enrichment and depletion of genic methylation classifications across different types of gene duplicates. Fisher exact test odds ratios of enriched associations are colored brown, depleted associations are in green. FDR corrected p-value < .05 are indicated in blue. Pagel's lambda test for phylogenetic signal is at the bottom of each table. A lambda value of '0' indicates no phylogenetic signal, while '1' indicates a strong phylogenetic signal. Pagel's lambda values are considered statistically significant with a FDR adjusted p-value < 0.05 and are highlighted in yellow.

Table S6: Number of duplicate gene pairs with different or the same genic methylation status for each type of duplication in each species.

Table S7: Number of pairs in each of the duplicate-pair methylation classification.

Table S8: Proportions of genes in each genic methylation class for the parental and daughter copies of translocated genes for each species.

Table S9: Median proportion across species of different methylation classes for parental and daughter translocated duplicates.

Table S10: Number of genes with similar or divergent methylation profiles between parental and translocated duplicates.

Table S11: Duplicate pair classifications ranked based on median Ks values for single gene duplicates (SGD) and whole-genome duplicates (WGD).

Table S12: Differences in the distribution of synonymous substitution rates (Ks) for duplicate gene pairs compared to a random distribution of the same number of paralogs. Blue indicates distribution is significantly different based on Kolmogorov-Smirnov test at an FDR adjusted p-value < 0.05.

Table S13: Outgroup species used for each epoch as part of MCscanX-transposed.

Table S14: Enrichment and depletion of genic methylation classifications across different epochs of transposed duplicates for all species. Fisher exact test odds ratios of enriched associations are colored green, depleted associations are in orange. Blue indicates distribution is significantly different at an FDR adjusted p-value < 0.05.

Table S15: Duplicate pair classifications ranked based on median Ka/Ks values for single gene duplicates (SGD) and whole-genome duplicates (WGD).

Table S16: Differences in the distribution of Ka/Ks ratios for duplicate gene pairs compared to a random distribution of the same number of paralogs. Blue indicates distribution is significantly different based on Kolmogorov-Smirnov test at an FDR adjusted p-value < 0.05.

Table S17: Enrichment and depletion of known presence-absence variants for gbM, teM, and unM genes. Fisher's Exact Test odds ratios of enriched associations are colored green, depleted associations are in orange. Blue indicates distribution is significantly different at a FDR adjusted p-value < 0.05.

Table S18: Correlations between genic methylation classes (gbM, teM, unM) and genomic features (number of genes, TEs, and TE nucleotides) in 100kb sliding windows with a 50kb step size. Positive correlations are marked in green, negative correlations in red. Blue indicates distribution is significantly different at a FDR adjusted p-value < 0.05.

Table S19: Pagel's lambda test for phylogenetic signal of correlations in Table S28. A lambda value of '0' indicates no phylogenetic signal, while '1' indicates a strong phylogenetic signal. Correlations in blue show a statistical significance of phylogenetic signal at FDR adjusted p < 0.05.

Table S20: Enrichment and depletion of transposable elements (TEs) with gbM, teM, and unM paralogs in each species. TEs within 3 kb upstream, downstream or within the gene body were considered associated with that gene. Fisher exact test odds ratio of less than 1 represents depletion (green), greater than 1 indicates enrichment (brown). Associations are considered significant at a FDR adjusted p-value < 0.05 and shown in blue.

Table S21: Enrichment and depletion of transposable elements (TEs) with different types of duplication in each species. TEs within 3 kb upstream, downstream or within the gene body were considered associated with that gene.. Fisher exact test odds ratio of less than 1 represents depletion (green), greater than 1 indicates enrichment (brown). Associations are considered significant at a FDR adjusted p-value < 0.05 and shown in blue.

Dataset S1: Phylogenetic tree of 43 angiosperms with branch lengths in newick format.

References:

1. L. E. Flagel, J. F. Wendel, Gene duplication and evolutionary novelty in plants. *New Phytologist* **183**, 557–564 (2009).

2. S. Ohno, Evolution by Gene Duplication (1970) https:/doi.org/10.1007/978-3-642-86659-3.

3. P. S. Soltis, D. B. Marchant, Y. Van de Peer, D. E. Soltis, Polyploidy and genome evolution in plants. *Curr. Opin. Genet. Dev.* **35**, 119–125 (2015).

4. S. P. Otto, J. Whitton, Polyploid Incidence and Evolution. *Annual Review of Genetics* **34**, 401–437 (2000).

5. F. Cheng, *et al.*, Gene retention, fractionation and subgenome differences in polyploid plants. *Nat Plants* **4**, 258–268 (2018).

6. Y. Van de Peer, E. Mizrachi, K. Marchal, The evolutionary significance of polyploidy. *Nat. Rev. Genet.* **18**, 411–424 (2017).

7. C. Simillion, K. Vandepoele, M. C. E. Van Montagu, M. Zabeau, Y. Van de Peer, The hidden duplication past of Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 13627–13632 (2002).

8. Y. Jiao, *et al.*, Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**, 97–100 (2011).

9. Amborella Genome Project, The Amborella genome and the evolution of flowering plants. *Science* **342**, 1241089 (2013).

10. X. Qiao, *et al.*, Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. *Genome Biol.* **20**, 38 (2019).

11. N. Jiang, Z. Bao, X. Zhang, S. R. Eddy, S. R. Wessler, Pack-MULE transposable elements mediate gene evolution in plants. *Nature* **431**, 569–573 (2004).

12. W. Wang, *et al.*, High rate of chimeric gene origination by retroposition in plant genomes. *Plant Cell* **18**, 1791–1802 (2006).

13. J. A. Bailey, *et al.*, Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).

14. J. Zhang, Evolution by gene duplication: an update. *Trends in Ecology & Evolution* **18**, 292–298 (2003).

15. M. Lynch, J. S. Conery, The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155 (2000).

16. S. Maere, *et al.*, Modeling gene and genome duplications in eukaryotes. *Proceedings of the National Academy of Sciences* **102**, 5454–5459 (2005).

17. M. Freeling, Bias in Plant Gene Content Following Different Sorts of Duplication: Tandem,

Whole-Genome, Segmental, or by Transposition. *Annual Review of Plant Biology* **60**, 433–453 (2009).

18. H. E. Cuevas, *et al.*, The Evolution of Photoperiod-Insensitive Flowering in Sorghum, A Genomic Model for Panicoid Grasses. *Mol. Biol. Evol.* **33**, 2417–2428 (2016).

19. X. Wang, *et al.*, Comparative genomic analysis of C4 photosynthetic pathway evolution in grasses. *Genome Biol.* **10**, R68 (2009).

20. K. Hanada, C. Zou, M. D. Lehti-Shiu, K. Shinozaki, S.-H. Shiu, Importance of Lineage-Specific Expansion of Plant Tandem Duplicates in the Adaptive Response to Environmental Stimuli. *Plant Physiology* **148**, 993–1003 (2008).

21. J. A. Fawcett, S. Maere, Y. Van de Peer, Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 5737–5742 (2009).

22. J. F. Wendel, The wondrous cycles of polyploidy in plants. *American Journal of Botany* **102**, 1753–1756 (2015).

23. J. C. Schnable, N. M. Springer, M. Freeling, Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 4069–4074 (2011).

24. R. De Smet, *et al.*, Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 2898–2903 (2013).

25. N. Panchy, M. D. Lehti-Shiu, S.-H. Shiu, Evolution of gene duplication in plants. *Plant Physiology*, 00523.2016 (2016).

26. S. N. Rodin, A. D. Riggs, Epigenetic silencing may aid evolution by gene duplication. *J. Mol. Evol.* **56**, 718–729 (2003).

27. H. Zhang, Z. Lang, J.-K. Zhu, Dynamics and function of DNA methylation in plants. *Nature Reviews Molecular Cell Biology* **19**, 489–506 (2018).

28. Y. Gruenbaum, T. Naveh-Many, H. Cedar, A. Razin, Sequence specificity of methylation in higher plant DNA. *Nature* **292**, 860–862 (1981).

29. P. Meyer, I. Niedenhof, M. ten Lohuis, Evidence for cytosine methylation of non-symmetrical sequences in transgenic Petunia hybrida. *EMBO J.* **13**, 2084–2088 (1994).

30. A. Zemach, I. E. McDaniel, P. Silva, D. Zilberman, Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328**, 916–919 (2010).

31. S. Feng, *et al.*, Conservation and divergence of methylation patterning in plants and animals. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 8689–8694 (2010).

32. S. K. K. Raju, S. K. Kenchanmane Raju, E. J. Ritter, C. E. Niederhuth, Establishment, maintenance, and biological roles of non-CG methylation in plants. *Essays in Biochemistry* **63**, 743–755 (2019).

33. X. Zhang, *et al.*, Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. *Cell* **126**, 1189–1201 (2006).

34. J. I. Gent, *et al.*, CHH islands: de novo DNA methylation in near-gene chromatin regulation in maize. *Genome Res.* **23**, 628–637 (2013).

35. C. E. Niederhuth, R. J. Schmitz, Putting DNA methylation in context: from genomes to gene expression in plants. *Biochim. Biophys. Acta Gene Regul. Mech.* **1860**, 149–156 (2017).

36. R. K. Tran, *et al.*, DNA methylation profiling identifies CG methylation clusters in Arabidopsis genes. *Curr. Biol.* **15**, 154–159 (2005).

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

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

39. S. Takuno, J.-H. Ran, B. S. Gaut, Evolutionary patterns of genic DNA methylation vary across land plants. *Nat Plants* **2**, 15222 (2016).

40. C. E. Niederhuth, *et al.*, Widespread natural variation of DNA methylation within angiosperms. *Genome Biol.* **17**, 194 (2016).

41. D. K. Seymour, D. Koenig, J. Hagmann, C. Becker, D. Weigel, Evolution of DNA Methylation Patterns in the Brassicaceae is Driven by Differences in Genome Organization. *PLoS Genetics* **10**, e1004785 (2014).

42. M. El Baidouri, *et al.*, Genic C-Methylation in Soybean Is Associated with Gene Paralogs Relocated to Transposable Element-Rich Pericentromeres. *Mol. Plant* **11**, 485–495 (2018).

43. K. L. Adams, R. Cronn, R. Percifield, J. F. Wendel, Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proceedings of the National Academy of Sciences* **100**, 4649–4654 (2003).

44. Z. Hua, *et al.*, Epigenomic programming contributes to the genomic drift evolution of the F-Box protein superfamily in Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 16927–16932 (2013).

45. T. E. Keller, S. V. Yi, DNA methylation and evolution of duplicate genes. *Proceedings of the National Academy of Sciences* **111**, 5932–5937 (2014).

46. A. Y.-F. Chang, B.-Y. Liao, DNA methylation rebalances gene dosage after mammalian

gene duplications. *Mol. Biol. Evol.* **29**, 133–144 (2012).

47. H. Wang, *et al.*, CG gene body DNA methylation changes and evolution of duplicated genes in cassava. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 13729–13734 (2015).

48. K. D. Kim, *et al.*, A Comparative Epigenomic Analysis of Polyploidy-Derived Genes in Soybean and Common Bean. *Plant Physiol.* **168**, 1433–1447 (2015).

49. L. Wang, *et al.*, Comparative epigenomics reveals evolution of duplicated genes in potato and tomato. *Plant J.* **93**, 460–471 (2018).

50. C. Xu, B. D. Nadon, K. D. Kim, S. A. Jackson, Genetic and epigenetic divergence of duplicate genes in two legume species. *Plant Cell Environ.* **41**, 2033–2044 (2018).

51. Y. Wang, X. Wang, T.-H. Lee, S. Mansoor, A. H. Paterson, Gene body methylation shows distinct patterns associated with different gene origins and duplication modes and has a heterogeneous relationship with gene expression inOryza sativa(rice). *New Phytologist* **198**, 274–283 (2013).

52. X. Wang, *et al.*, Gene-body CG methylation and divergent expression of duplicate genes in rice. *Scientific Reports* **7** (2017).

53. J. Wang, N. C. Marowsky, C. Fan, Divergence of gene body DNA methylation and evolution of plant duplicate genes. *PLoS One* **9**, e110357 (2014).

54. J. F. Wendel, D. Lisch, G. Hu, A. S. Mason, The long and short of doubling down: polyploidy, epigenetics, and the temporal dynamics of genome fractionation. *Current Opinion in Genetics & Development* **49**, 1–7 (2018).

55. Z. Li, *et al.*, Gene Duplicability of Core Genes Is Highly Consistent across All Angiosperms. *Plant Cell* **28**, 326–344 (2016).

56. A. H. Paterson, *et al.*, Many gene and domain families have convergent fates following independent whole-genome duplication events in Arabidopsis, Oryza, Saccharomyces and Tetraodon. *Trends Genet.* **22**, 597–602 (2006).

57. A. J. Bewick, *et al.*, On the origin and evolutionary consequences of gene body DNA methylation. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 9111–9116 (2016).

58. Y. Wang, J. Li, A. H. Paterson, MCScanX-transposed: detecting transposed gene duplications based on multiple colinearity scans. *Bioinformatics* **29**, 1458–1460 (2013).

59. X. P. Zhao, *et al.*, Dispersed repetitive DNA has spread to new genomes since polyploid formation in cotton. *Genome Res.* **8**, 479–492 (1998).

60. M. Freeling, *et al.*, Many or most genes in Arabidopsis transposed after the origin of the order Brassicales. *Genome Research* **18**, 1924–1937 (2008).

61. B. P. Cusack, K. H. Wolfe, Not born equal: increased rate asymmetry in relocated and retrotransposed rodent gene duplicates. *Mol. Biol. Evol.* **24**, 679–686 (2007).

62. E. W. Ganko, B. C. Meyers, T. J. Vision, Divergence in expression between duplicated genes in Arabidopsis. *Mol. Biol. Evol.* **24**, 2298–2309 (2007).

63. T. P. G. S. Consortium, The Potato Genome Sequencing Consortium, Genome sequence and analysis of the tuber crop potato. *Nature* **475**, 189–195 (2011).

64. T. Münkemüller, *et al.*, How to measure and test phylogenetic signal. *Methods in Ecology and Evolution* **3**, 743–756 (2012).

65. M. Pagel, Inferring the historical patterns of biological evolution. *Nature* **401**, 877–884 (1999).

66. J. C. Dohm, *et al.*, Palaeohexaploid ancestry for Caryophyllales inferred from extensive gene-based physical and genetic mapping of the sugar beet genome (Beta vulgaris). *Plant J.* **70**, 528–540 (2012).

67. J. C. Dohm, *et al.*, The genome of the recently domesticated crop plant sugar beet (Beta vulgaris). *Nature* **505**, 546–549 (2014).

68. S. Tasdighian, *et al.*, Reciprocally Retained Genes in the Angiosperm Lineage Show the Hallmarks of Dosage Balance Sensitivity. *Plant Cell* **29**, 2766–2785 (2017).

69. A. V. Klepikova, A. S. Kasianov, E. S. Gerasimov, M. D. Logacheva, A. A. Penin, A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. *Plant J.* **88**, 1058–1070 (2016).

70. R. F. McCormick, *et al.*, The Sorghum bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *Plant J.* **93**, 338–354 (2018).

71. J. Wang, *et al.*, SoyCSN: Soybean context-specific network analysis and prediction based on tissue-specific transcriptome data. *Plant Direct* **3** (2019).

72. I. Yanai, *et al.*, Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).

73. S. J. Cokus, *et al.*, Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **452**, 215–219 (2008).

74. R. Lister, *et al.*, Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133**, 523–536 (2008).

75. R. J. Schmitz, *et al.*, Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. *Genome Res.* **23**, 1663–1674 (2013).

76. T. Kawakatsu, *et al.*, Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. *Cell* **166**, 492–505 (2016).

77. M. W. Vaughn, *et al.*, Epigenetic Natural Variation in Arabidopsis thaliana. *PLoS Biology* **5**, e174 (2007).

78. S. R. Eichten, *et al.*, Epigenetic and genetic influences on DNA methylation variation in maize populations. *Plant Cell* **25**, 2783–2797 (2013).

79. S. R. Eichten, T. Stuart, A. Srivastava, R. Lister, J. O. Borevitz, DNA methylation profiles of diverse Brachypodium distachyon align with underlying genetic diversity. *Genome Res.* **26**, 1520–1531 (2016).

80. P. P. Edger, *et al.*, Subgenome Dominance in an Interspecific Hybrid, Synthetic Allopolyploid, and a 140-Year-Old Naturally Established Neo-Allopolyploid Monkeyflower. *The Plant Cell* **29**, 2150–2167 (2017).

81. K. A. Bird, *et al.*, Replaying the evolutionary tape to investigate subgenome dominance in allopolyploid Brassica napus. *New Phytol.* **230**, 354–371 (2021).

82. M. A. Matzke, R. A. Mosher, RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat. Rev. Genet.* **15**, 394–408 (2014).

83. J. Bender, G. R. Fink, Epigenetic control of an endogenous gene family is revealed by a novel blue fluorescent mutant of Arabidopsis. *Cell* **83**, 725–734 (1995).

84. R. A. Veitia, S. Bottani, J. A. Birchler, Cellular reactions to gene dosage imbalance: genomic, transcriptomic and proteomic effects. *Trends Genet.* **24**, 390–397 (2008).

85. F. A. Kondrashov, Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proc. Biol. Sci.* **279**, 5048–5057 (2012).

86. S. D. Pophaly, A. Tellier, Population Level Purifying Selection and Gene Expression Shape Subgenome Evolution in Maize. *Mol. Biol. Evol.* **32**, 3226–3235 (2015).

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

88. J. L. Bennetzen, C. Coleman, R. Liu, J. Ma, W. Ramakrishna, Consistent over-estimation of gene number in complex plant genomes. *Curr. Opin. Plant Biol.* **7**, 732–736 (2004).

89. J. C. Schnable, Genes and gene models, an important distinction. *New Phytol.* (2019) https:/doi.org/10.1111/nph.16011.

90. A. B. Silveira, *et al.*, Extensive Natural Epigenetic Variation at a De Novo Originated Gene. *PLoS Genetics* **9**, e1003437 (2013).

91. X. Jiang, R. Assis, Rapid functional divergence after small-scale gene duplication in grasses. *BMC Evol. Biol.* **19**, 97 (2019).

92. S.-S. C. Huang, J. R. Ecker, Piecing together cis-regulatory networks: insights from epigenomics studies in plants. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **10**, e1411 (2018).

93. M. D. Schultz, *et al.*, Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* **523**, 212–216 (2015).

94. Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).

95. M. D. Schultz, R. J. Schmitz, J. R. Ecker, "Leveling" the playing field for analyses of single-base resolution DNA methylomes. *Trends in Genetics* **28**, 583–585 (2012).

96. D. M. Emms, S. Kelly, OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).

97. D. M. Emms, S. Kelly, OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).

98. B. Buchfink, C. Xie, D. H. Huson, Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**, 59–60 (2015).

99. S. R. A. Fisher, *Statistical Methods for Research Workers* (1934).

100. , Various R Programming Tools for Plotting Data [R package gplots version 3.1.1] (2020) (June 5, 2021).

101. Y. Jin, H. Qian, V.PhyloMaker: an R package that can generate very large phylogenies for vascular plants. *Ecography* **42**, 1353–1359 (2019).

102. L. J. Revell, phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* **3**, 217–223 (2012).

103. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

104. M. Suyama, D. Torrents, P. Bork, PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–12 (2006).

105. D. Wang, Y. Zhang, Z. Zhang, J. Zhu, J. Yu, KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinformatics* **8**, 77–80 (2010).

106. F. J. Massey, The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association* **46**, 68–78 (1951).

107. A. A. Golicz, *et al.*, The pangenome of an agronomically important crop plant Brassica oleracea. *Nat. Commun.* **7**, 13390 (2016).

108. L. Gao, *et al.*, The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet.* **51**, 1044–1051 (2019).

109. M. A. Hardigan, *et al.*, Genome Reduction Uncovers a Large Dispensable Genome and Adaptive Role for Copy Number Variation in Asexually Propagated Solanum tuberosum. *Plant Cell* **28**, 388–405 (2016).

110. C. N. Hirsch, *et al.*, Insights into the maize pan-genome and pan-transcriptome. *Plant Cell* **26**, 121–135 (2014).

111. J. A. O'Rourke, *et al.*, An RNA-Seq based gene expression atlas of the common bean. *BMC Genomics* **15**, 866 (2014).

112. A. Dobin, *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

113. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

114. D. M. Goodstein, *et al.*, Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, D1178–86 (2012).

115. S. Ou, *et al.*, Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).

116. T. Wei, S. Viliam, R package "corrplot": Visualization of a Correlation Matrix (Version 0.84) (June 5, 2021).

117. J. Pellicer, I. J. Leitch, The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. *New Phytologist* **226**, 301–305 (2020).

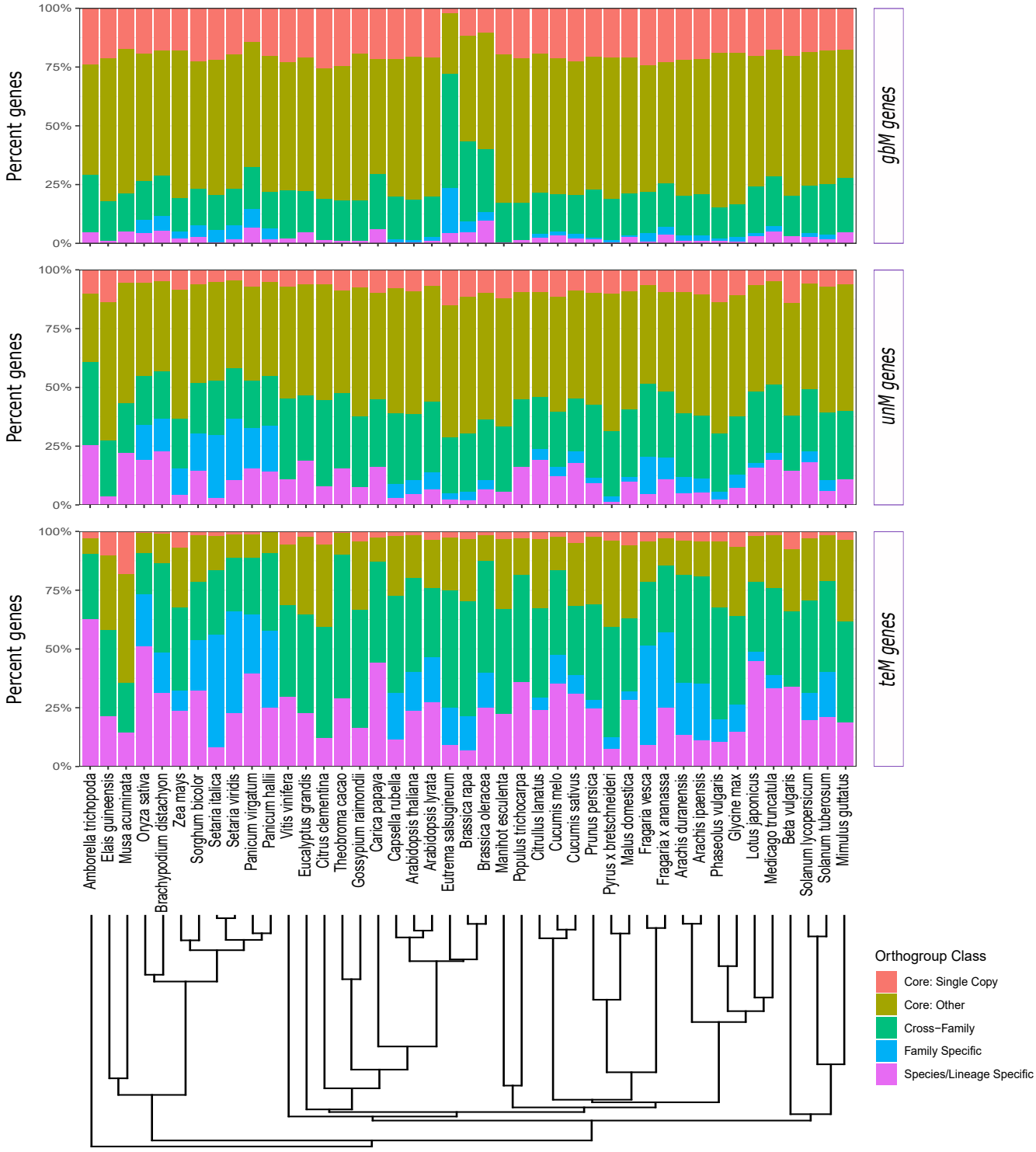

Figure 1: Distribution of orthogroups in genic methylation classes. For each species, the percentage of genes classified into different orthogroup categories (core-single copy, core-other, cross-family, family-specific, and species/lineage-specific) in each of the three genic methylation classification (gbM, teM, and unM genes).

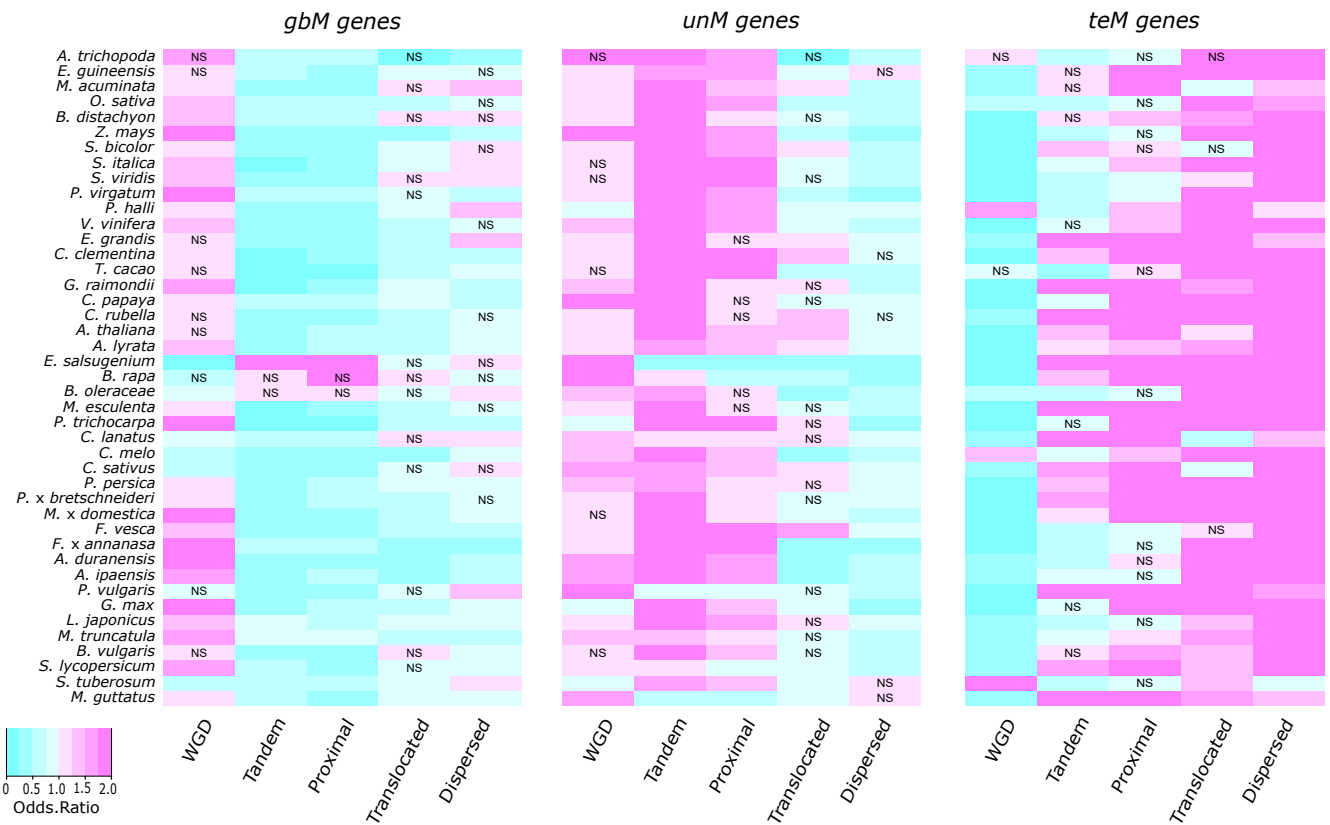

Figure 2: Patterns of genic methylation across different types of gene duplicates. Enrichment or depletion of each genic methylation class (gbM, teM, and unM) for each type of gene duplication (WGD, tandem, proximal, translocated, and dispersed). Increasing shades of cyan indicates greater depletion, while increasing shades of magenta represents greater enrichment. Unless indicated, all associations are statistically significant at a FDR-corrected p-value < 0.05. 'NS' indicates no statistical significance.

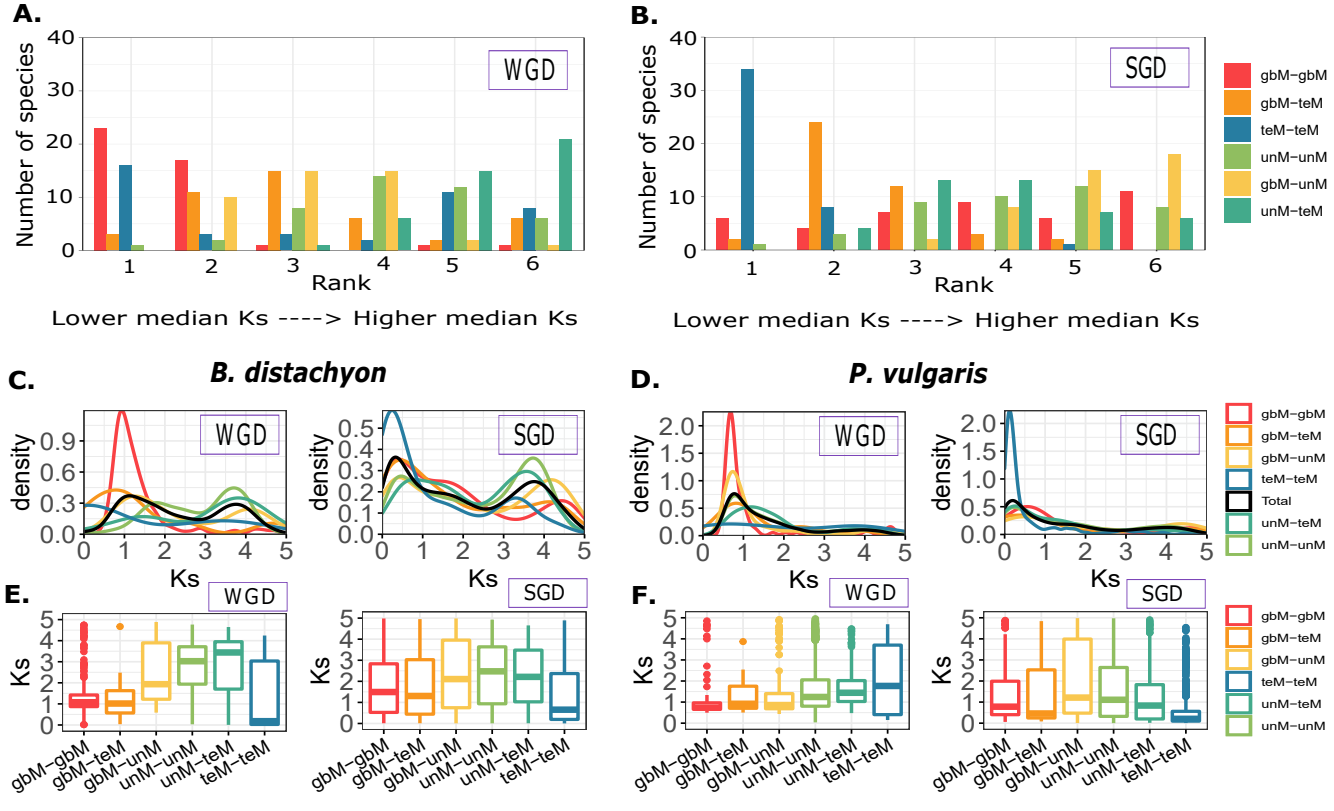

Figure 3: Relationship between genic methylation and the age of gene duplication. Bar plots showing the number of species in each of the duplicate-pair genic methylation classifications (gbM-gbM, gbM-teM, teM-teM, unM-unM, gbM-unM, and unM-teM) ranked based on median Ks values (synonymous substitutions) for whole-genome duplicates (A) and single-gene duplicates (B). Density plots (C and D) and box plots (E and F) show the distribution of synonymous substitutions (Ks) for each of the duplicate-pair genic methylation classificationsin *B. distachyon* and *P. vulgaris* respectively. The black line in the density plots represents the Ks distribution of all duplicate pairs.

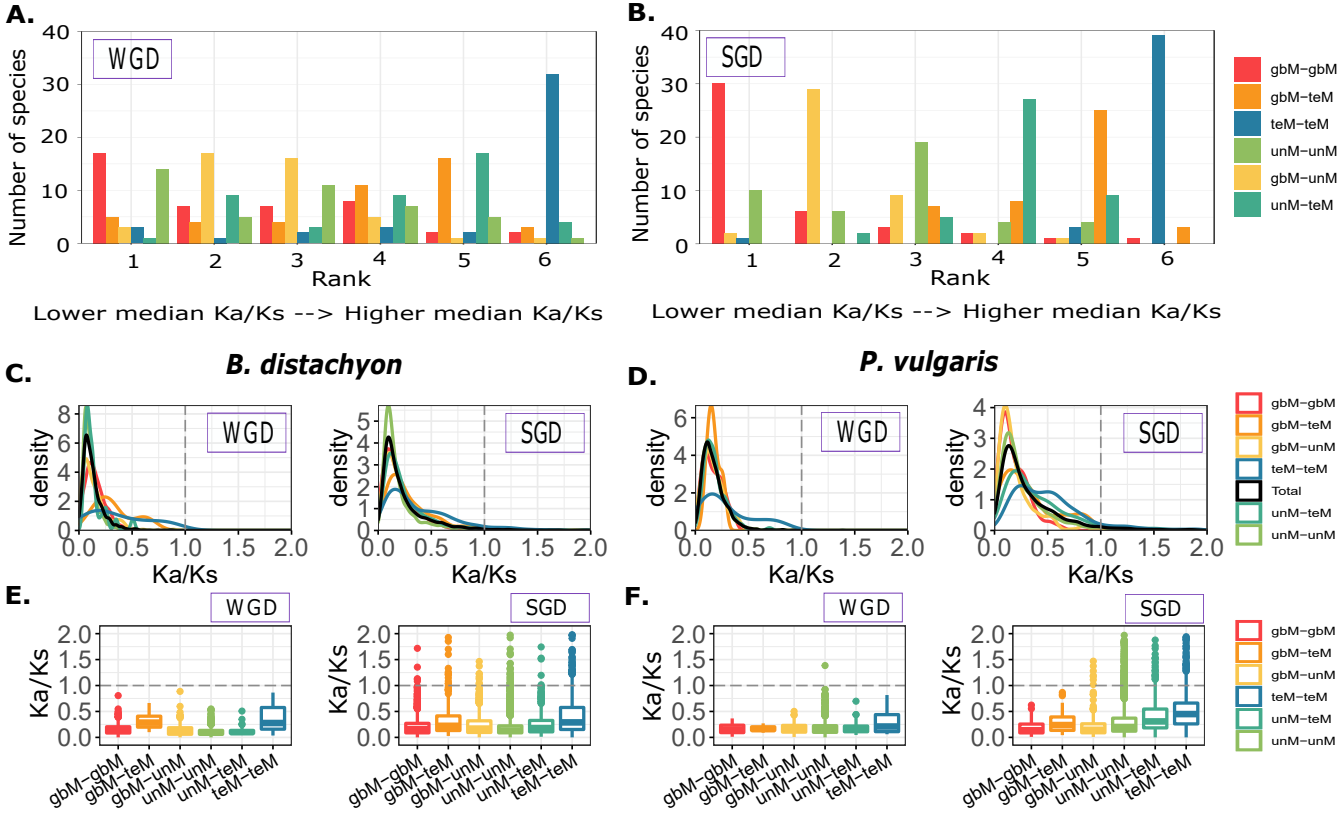

Figure 4: Relationship between genic methylation and sequence evolution for duplicate pairs. Bar plots showing the number of species in each of the duplicate-pair genic methylation classifications (gbM-gbM, gbM-teM, teM-teM, unM-unM, gbM-unM, and unM-teM) ranked based on median Ka/Ks values (ratio of Ka, non-synonymous substitutions to Ks, synonymous substitutions) for whole-genome (A) and single-gene duplicates (B). Density plots (C and D) and box plots (E and F) show the distribution of Ka/Ks ratios for each of the duplicate-pair genic methylation classifications in B. distachyon and P. vulgaris respectively. Dotted line at Ka/Ks ratio of '1' suggestive of neutral selection. Black line in the density plots represents the Ka/Ks distribution of all duplicate pairs.

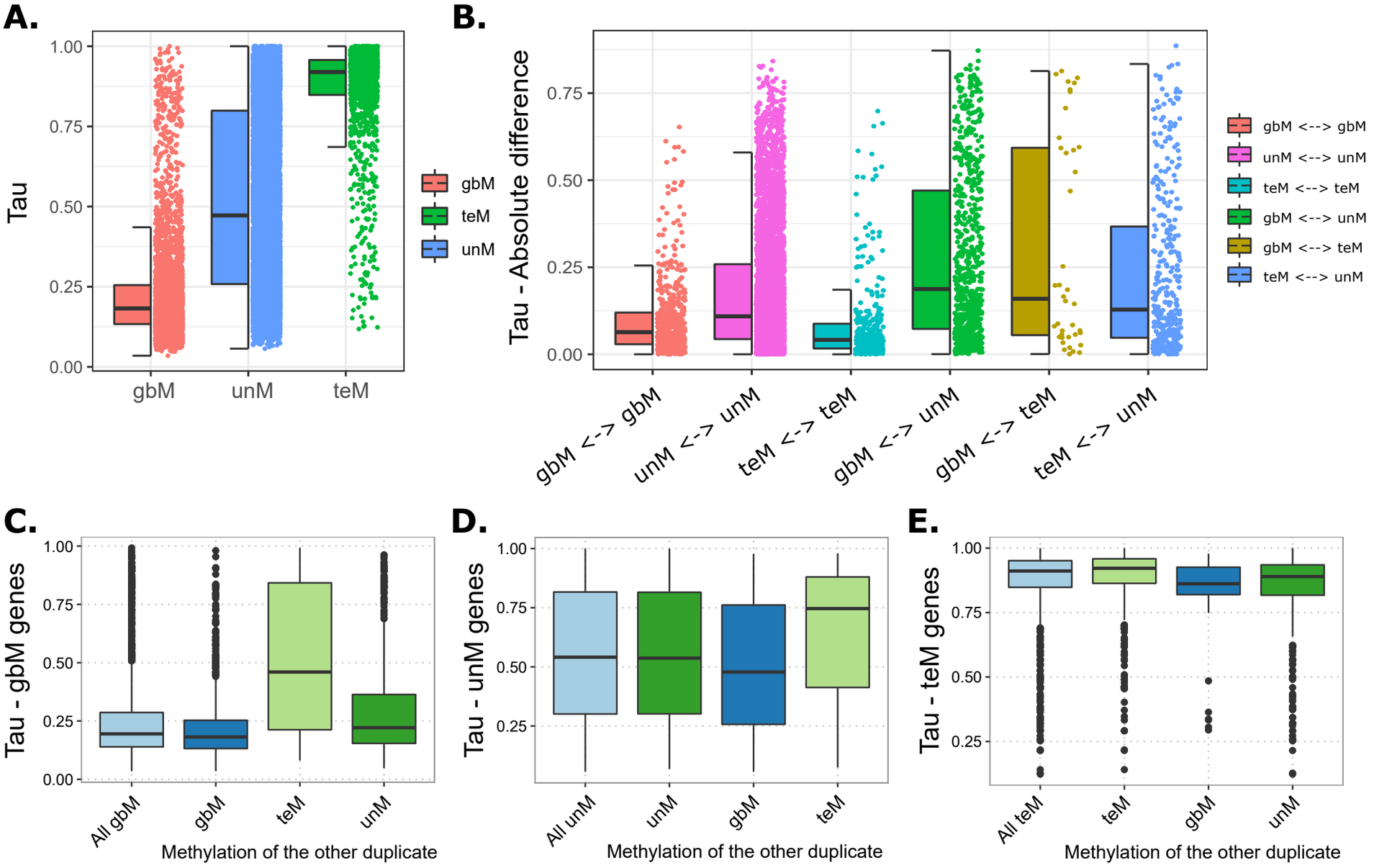

Figure 5: Gene expression specificity of A. thaliana duplicate gene pairs. Tissue-specificity index, Tau (τ), ranges from 0 (broadly expressed) to 1 (narrowly expressed). (A) Tissue specificity of genes based on genic methylation classification (gbM, unM, and teM). (B) Absolute difference in tissue -specificity index (τ) between pairs of duplicate genes with similar or divergent methylation. Differences in Tau specificity of gbM, unM, and teM genes (C, D, and E respectively) when the other duplicate pair has the same or a different genic methylation status. For example, for gbM genes, the tau specificity was plotted for all gbM genes and the gbM paralog in gbM-gbM, gbM-teM, and gbM-unM pairs. For unM genes, the tau of only the unM paralog is shown and similarly for teM genes, only the tau of the teM paralog is shown.

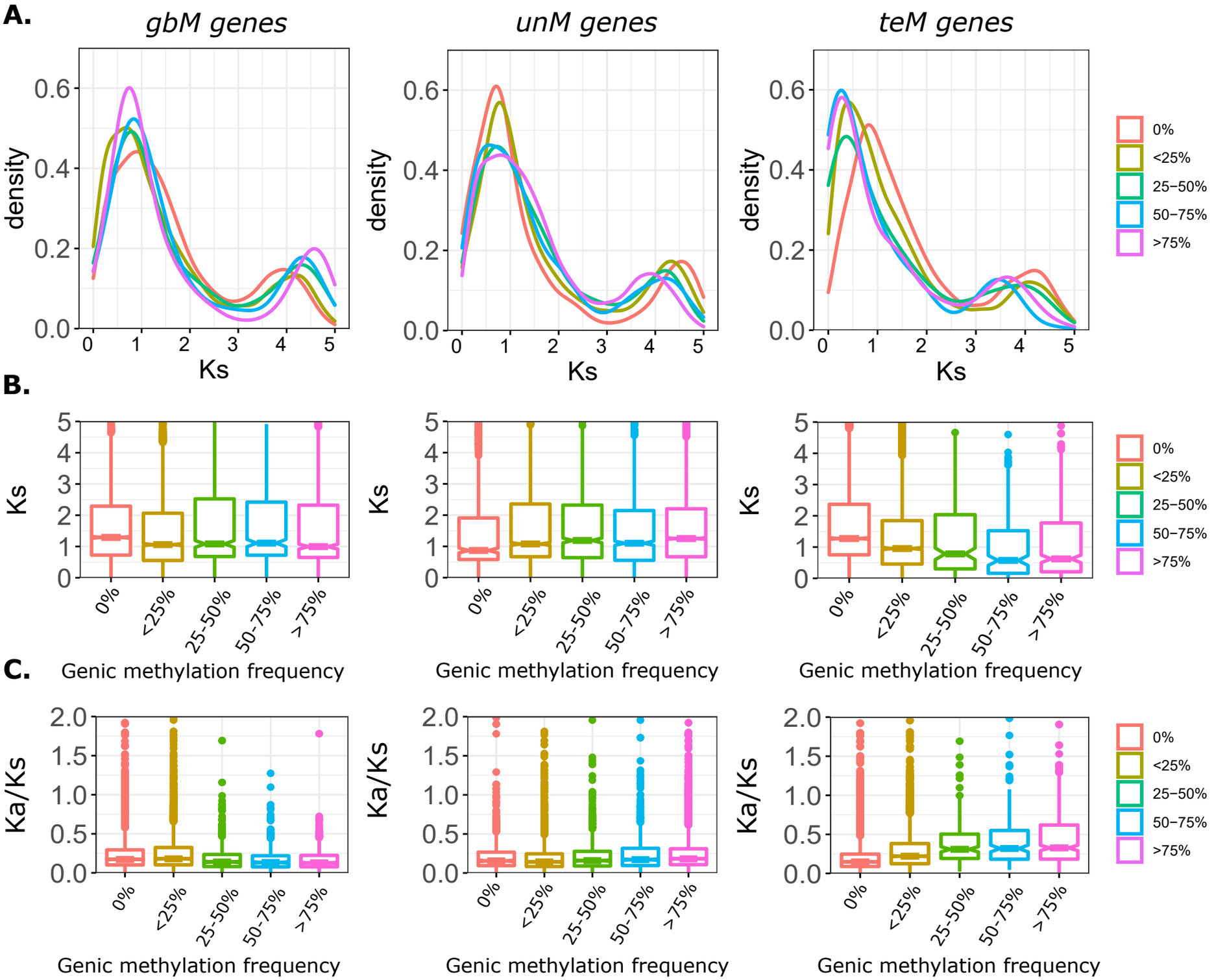

Figure 6: Association of genic methylation frequency with theage of gene duplication and sequence evolution in 928 *A. thaliana* accessions. A) Density plots showing the Ks distributionof genes at different frequencies (0%, <25%, 25%-50%, 50%-75%, >75%) in the population for gbM, unM, and teM genes. Boxplots of Ks (B) and Ka/Ks distributions (C) for gbM, unM, and teM genes at different frequencies.