# 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 an ongoing source of genetic novelty and evolutionary innovation. DNA methylation has been proposed as a factor in duplicate gene evolution. However, previous studies have largely been limited to individual species, apply differing methods, and have predominantly focused on CG methylation. The generalizability of these relationships at phylogenetic and population scales remains unknown. Here, we apply a consistent comparative epigenomics approach across 43 diverse angiosperm species and a population of 928 *Arabidopsis thaliana* accessions examining both CG and non-CG methylation contexts n whole-genome duplicate (WGD) and single-gene duplicate (SGD) genes. We observe several overarching trends, demonstrating that genic DNA methylation is differentially associated with the type of gene duplication, age of gene duplication, sequence evolution, and gene expression. WGDs are typically unmethylated or marked by mCG-only gene-body methylation, while SGDs typically are enriched for unmethylated genes or methylation in both mCG and non-CG contexts (transposon-like methylation or teM). TeM in particular was associated with relatively more recent SGDs and higher sequence divergence. However, we find variation across the phylogeny, as well as phylogenetic effects, which could be key to a deeper understanding of these relationships. Within the *A. thaliana* population, differences in duplicate age and sequence evolution were observed based on the frequency of genic methylation. Collectively, these results indicate that genic methylation states differently mark duplicate gene evolution.

**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 (Flagel and Wendel 2009; Ohno 1970; Bridges 1935). Whole-genome duplication (WGD), or polyploidy, results in an increase in the entire genomic content of an organism (Soltis et al. 2015) and is more pervasive in plants than other eukaryotic lineages (Otto and Whitton 2000; F. Cheng et al. 2018; Van de Peer, Mizrachi, and Marchal 2017). Small-scale and single-gene duplications (SGDs) also significantly contribute to the gene repertoire (Panchy, Lehti-Shiu, and Shiu 2016). SGD is a continuous process, with ongoing gene birth and death (Lynch and Conery 2000; Maere et al. 2005). The subsequent retention, divergence, or loss of paralogs (duplicated genes) is biased depending on the type of duplication and gene function

(Michael Freeling 2009; De Smet et al. 2013). The factors determining the evolutionary fate of paralogs remains an area of intense study and DNA methylation is thought to be a contributing factor due to its influence on gene expression (Rodin and Riggs 2003; Y. Wang et al. 2013).

Cytosine Methylation at CG dinucleotides is found throughout plants, animals, and fungi, while methylation of the non-CG trinucleotide CHG and CHH (H = A, T or C) contexts is limited to plants (Zemach et al. 2010; Feng et al. 2010). Non-CG methylation is predominantly found in heterochromatin alongside CG methylation and is involved in TE silencing (Raju et al. 2019). Plant genes have several distinct patterns of DNA methylation within coding regions (henceforth 'genic methylation') with strong associations with gene expression (Niederhuth and Schmitz 2017). Genes characterized by CG-only methylation in coding regions are referred to as gene-body methylated (gbM) (Tran et al. 2005; X. Zhang et al. 2006). GbM is usually conserved between orthologous genes and gbM genes typically have broad expression and evolve more slowly (Takuno and Gaut 2013, 2012; Takuno, Ran, and Gaut 2016; Niederhuth et al. 2016). Some genes are methylated similar to transposable-elements (TEs), having both CG and non-CG methylation within coding regions. This transposable-element like methylation (teM) is rarely conserved between orthologs and associated with transcriptional silencing and narrow expression patterns (Niederhuth et al. 2016; Seymour et al. 2014; El Baidouri et al. 2018). The majority of genes, however, are unmethylated (unM) and have variable expression across tissues and conditions (Takuno and Gaut 2012; Niederhuth et al. 2016).

DNA methylation could serve to buffer the genome against changes in gene dosage by modulating gene expression and facilitating functional divergence. For instance, in Arabidopsis, DNA methylation was associated with divergence of the highly duplicated F-box gene family (Hua et al. 2013). Silencing by DNA methylation could result in either retention or loss of paralogs. Tissue-specific silencing of paralogs could lead to sub-functionalization of paralog expression and retention (Adams et al. 2003; Rodin and Riggs 2003). Alternatively, silencing may contribute to pseudogenization and subsequent paralog loss (El Baidouri et al. 2018; Hua et al. 2013). In animals, DNA methylation differences between paralogs correspond to divergence in gene expression, with more recent duplicate genes being more heavily methylated (Keller and Yi 2014; Chang and Liao 2012). Associations have been found between gbM and the divergence of CG methylation with gene expression and sequence evolution in different plant species (C. Xu et al. 2018; H. Wang et al. 2015; X. Wang et al. 2017; Jun Wang, Marowsky, and Fan 2014; Kim et al. 2015; L. Wang et al. 2018; Y. Wang et al. 2013). In soybeans, non-CG methylation was found to associate with paralogs transposed to heterochromatic regions and with increased sequence divergence (El Baidouri et al. 2018).

While past studies have explored the relationship between DNA methylation and gene duplication, these have been limited to individual species and typically focused only on CG methylation. Lineage-specific variation in DNA methylation (Niederhuth et al. 2016), histories of gene duplication (Qiao et al. 2019), and differences in analysis have precluded a general understanding of this relationship. To address these issues, we analyze DNA methylation across 43 angiosperm species and a population of 928 *Arabidopsis thaliana* ecotypes across all genic methylation contexts. We identify overarching trends and relationships between genic methylation, the type and age of duplication, and paralog evolution. This work provides a broad phylogenetic and population scale understanding of the role of DNA methylation in the evolution of plant duplicate genes.

# Results:

## Patterns of genic methylation across gene duplicates

DNA methylation (Niederhuth et al. 2016), gene content (Z. Li et al. 2016), and the extent of gene duplication (Qiao et al. 2019) vary across angiosperms. We first asked if there were general associations between these factors at a phylogenetic level. Genes from 43 angiosperm species with available whole-genome bisulfite sequencing data (Table S1) were classified based on genic methylation as either gene-body methylated (gbM), unmethylated (unM), or transposable-element like methylated (teM) (Table S2). To assess gene content, we then identified orthogroups, gene sets derived from a common ancestor, for these 43 species and an additional 15 angiosperm species to serve as outgroups (Table S1). As previously observed (Z. Li et al. 2016), orthogroups are bimodally distributed (Figure S1). Orthogroups found in ≥85% of species were classified as 'core angiosperm' genes and further subdivided as either 'core-single copy', if represented by a single gene in ≥ 70% of species, or 'core-other'. Remaining orthogroups were then 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.

We then examined the distribution of genic methylation across orthogroup categories (Figure S2, Table S3). GbM genes are predominantly found in core angiosperm orthogroups, with 2.1-25.6% (median – 20%) in core-single-copy and 25.5-65.7% (median – 57%) in core-others. Representation of gbM decreased with increasing lineage specificity. There were three exceptions to this trend: *Eutrema salsugineum*, *Brassica rapa*, and *Brassica oleracea*. These three species are known to be highly depleted of gbM, with near complete loss of gbM in *E. salsugineum* (Bewick et al. 2016). A large proportion of unM genes are also from core angiosperm orthogroups, with 4.6-15.3% (median – 9%) core-single-copy and 29-58.6% (median – 47%) core-other. However, a higher proportion of unM genes are in non-core angiosperm orthogroups than gbM genes. In contrast teM genes have comparatively fewer core angiosperm orthogroups, with single-copy orthogroups (0.5-18.4%, median – 3%) and core-other (6.5-46%, median – 22.3%). Instead, they are found in greater proportions in more lineage-specific orthogroups: cross-family (17.6-60.8%, median – 38%), family-specific (0-47.8%, median – 9.5%), and lineage/species-specific orthogroups (6.9-62.6%, median – 24%). This suggests gbM and unM are much more highly conserved and that teM is biased towards less conserved genes of more recent evolutionary origin.

Duplicated genes were next identified and classified (Table S4) as either whole-genome duplicates (WGDs) or one of four types of single-gene duplicates (SGDs). Tandem duplicates are thought to occur through unequal crossing-over, creating clusters of two or more genes adjacent to each other (J. Zhang 2003). Proximal duplicates are separated by several intervening genes and arose either through local transposition or interruption of an ancient tandem duplication (Zhao et al. 1998; M. Freeling et al. 2008). Translocated duplicates (also known as 'transposed') are pairs in which one of the genes is syntenic and the other is non-syntenic (Qiao et al. 2019; Y. Wang, Li, and Paterson 2013). Translocated duplicates can arise either by retrotransposition or DNA-based duplication (Cusack and Wolfe 2007) and the syntenic gene is assumed to be the

parental copy (Y. Wang, Li, and Paterson 2013). Finally, dispersed duplicates are pairs that fit none of the above criteria and can arise through multiple mechanisms (Michael Freeling 2009; Ganko, Meyers, and Vision 2007; Qiao et al. 2019).

Each class of gene duplication was tested for enrichment or depletion of gbM, unM, or teM (Figure 1, Table S5) revealing a number of general trends. WGDs typically are enriched for gbM (28/43 enriched, 6/43 depleted) and unM (33/43 enriched, 4/43 depleted), and depleted in teM (3/43 enriched, 38/43 depleted). Amongst the four classes of SGDs, we observe a broader trend based on whether the duplication is 'local' (tandem and proximal) or 'distal' (translocated and dispersed). Local SGDs were depleted of gbM (40/43 depleted, 1/43 enriched) in all species except for the three gbM-deficient Brassicaceae species, and enriched for unM in the majority of species (tandem – 40/43 enriched, 3/43 depleted; proximal – 33/43 enriched, 5/43 depleted). TeM was highly variable in tandem duplicates (18/43 enriched, 18/43 depleted), while proximal duplicates show more teM (29/43 enriched, 3/43 depleted). Distal SGDs are generally enriched for teM (translocated – 37/43 enriched, 3/43 depleted; dispersed – 42/43 enriched, 1/43 depleted) and depleted in gbM (translocated – 0/43 enriched, 30/43 depleted; dispersed – 9/43 enriched, 23/43 depleted) and unM (translocated – 9/43 enriched, 19/43 depleted; dispersed – 0/43 enriched, 38/43 depleted). The increasing enrichment of teM from tandem to proximal to distal SGDs suggests that teM becomes more common as genes move to increasingly different sequence or chromatin environments.

While this comparison revealed multiple trends across types of gene duplication, exceptions were found to every trend. To better understand these exceptions and reveal possible phylogenetic patterns, we tested these patterns of enrichment and depletion for phylogenetic signal using Pagel's lambda (λ) (Münkemüller et al. 2012; Pagel 1999), which ranges from 0 (no phylogenetic signal) to 1 (strong phylogenetic signal). Phylogenetic signal was 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 gbM-deficient Brassicaceae species (λ ~ 0.79). All three of the Cucurbitaceae species in our data are depleted for gbM in WGDs, which may be driving this result. Another notable exception was *Solanum tuberosum*, which is the only species where WGDs show the exact opposite pattern, being depleted in gbM and unM and enriched in teM. It is also the only species depleted for teM in dispersed duplicates. *S. tuberosum* is an autotetraploid and its last WGD is relatively recent compared to the other species in these data (Consortium and The Potato Genome Sequencing Consortium 2011; L. Wang et al. 2018), both of which may be factors and require further investigation.

**Frequency of genic methylation switching between paralogs**

As differences in genic methylation between paralogs could facilitate divergence, we next assessed the degree to which paralogs shared the same or different genic methylation profiles (Figure 2, Table S6, S7). WGDs had the highest similarity across species (same: ~70-97%, median – 85%; different: ~2-30%, median – 15%), followed by tandem (same: ~69%-93%, median – 82%; different: ~7-31%, median – 18%), proximal (same: ~66%-90%, median – 78%; different: ~10-34%, median – 22%), and dispersed (same: ~65%-92%, median – 77%; different: ~8-35%, median – 23%). Translocated duplicates had the broadest range and the lowest proportion of paralogs with similar genic methylation across species (same: ~51%-91%, median – 75%; different: ~9-48%, median – 25%). In cases where paralogs differ in genic methylation,

we typically cannot discern the direction of the change. However, for translocated duplicates, one of the paralogs is syntenic and considered the parental locus (Y. Wang, Li, and Paterson 2013). Assuming that methylation at the parental locus is the original state, we can determine the directionality of genic methylation changes in the translocated copy. The translocated copy had a higher proportion of teM in 33/43 species and a lower proportion of gbM in 26/43 species, while the translocated copy had a higher proportion of unM in 4/43 species and a lower proportion in 8/43 species (Table S8). A more specific examination of the direction of genic methylation switching shows that switching to teM was the most common form in 20/43 species, while switching to unM was more common in 20/43 species and to gbM in 3/43 species (Table S9). These results indicate that regardless of duplication type, the majority of paralogs have similar genic methylation states. However, differences in genic methylation become more common as duplicate copies are placed in distant genomics regions from each other. The directionality of this switching, at least for translocated copies, can vary and is not uniformly towards silencing by teM.

**Gene duplicate age associates with genic methylation**

Synonymous substitutions (Ks) are assumed to accumulate neutrally with time and Ks distributions have been widely used to date gene duplication events (Lynch and Conery 2000; Maere et al. 2005). We determined pairwise Ks values for all duplicate pairs, which returned a single value that applies to both genes in that pair. To avoid double counting, we only examined Ks distributions in the context of duplicate pairs, as opposed to individual genes. Duplicate pairs were put into six groups based on the genic methylation of each gene in the pair, e.g., a 'gbM-unM' has one gbM paralog and one unM paralog. Ks distributions were then examined for WGDs and SGDs to determine the relative age of each of these groups of duplicate pairs (Figure 3; Figure S3; Table S10, S11). All four types of SGDs (tandem, proximal, translocated, and dispersed) showed similar Ks distributions and were therefore examined together.

To compare trends across the phylogeny, we ordered duplicate pair groups, lowest to highest, based on median Ks (Figure 3A,B, Table S10). WGD pairs in which one or both genes are gbM typically have lower median Ks values, especially gbM-gbM pairs; while pairs containing unM genes, in particular unM-teM or unM-unM pairs, typically have a higher median Ks. We did not observe any trend for WGD teM-containing pairs. While there is a group of species in which teM-teM pairs have the lowest median Ks, this may be a spurious result due to the scarcity of WGD teM-teM pairs. At face-value, these results suggest that gbM-containing WGD pairs are typically younger, while unM-containing WGD pairs are older. However, these results must be considered within the context of the history of WGD and other factors (see Discussion). WGD results in the duplication of all genes and for many species, the last WGD event was millions of years ago. For example, the last WGD event in the *Beta vulgaris* genome was the core-eudicot γ WGT(Qiao et al. 2019; Dohm et al. 2012, 2014), therefore the comparatively younger age of many WGD gbM-gbM pairs can still reflect millions of years of retention.

In contrast to WGD, SGD is a continuous process. For SGDs, teM-containing pairs typically have lower median Ks values. This is most evident amongst teM-teM pairs, but gbM-teM and unM-teM show this same trend. GbM- and unM-containing pairs have higher median Ks than teM-containing pairs, but otherwise show no obvious trend (Figure 3B). This suggests that teM SGD genes are evolutionarily younger and more recent in origin. However, there are notable

exceptions. In both *Medicago truncatula* and *Pyrus x. bretschneideri* teM SGD pairs have higher median Ks values. We further tested this relationship for translocated genes using a method independent of the Ks-based approach. As the syntenic gene is assumed to be parental, the daughter gene can be parsed into different periods (epochs) since speciation, by sequential exclusion to the closest outgroup (Table S12), as employed in the program *MCScanX-transposed* (Y. Wang, Li, and Paterson 2013). More recent translocated duplicates were enriched in teM genes, while more ancient translocated duplicates were enriched for gbM and unM genes (Figure S4, Table S13), confirming our results from the Ks analysis.

**Genic methylation marks differences in paralog sequence evolution**

The ratio of non-synonymous (Ka) to synonymous (Ks) substitutions (Ka/Ks) is indicative of selection; with Ka/Ks < 1 indicative of purifying selection, Ka/Ks = 0 of neutral selection, and Ka/Ks > 1 indicative of diversifying selection. We determined Ka/Ks ratios for each duplicate pair and examined their distributions as was done above for Ks distributions (Figure 4, Figure S5, Table S14, S15). The vast majority of duplicate pairs have a Ka/Ks < 1, regardless of the type of duplication or genic methylation. However, there are differences in the distribution based on genic methylation. For both WGD and SGD genes, teM-containing pairs, in particular teM-teM pairs, have higher median Ka/Ks values; while gbM-containing pairs, especially gbM-gbM pairs, have lower median Ka/Ks. This suggests that despite Ka/Ks < 1, teM genes may be under relaxed selective constraints compared to gbM and unM. Interestingly, teM-containing pairs in many species are enriched for Ka/Ks > 1 compared to unM or gbM containing pairs indicating greater diversifying selection, although this needs more rigorous testing to confirm (Figure S5, Table S16).

Ongoing gene duplication and loss within a population will create copy number and presence absence variation (PAV). We examined the relationship between genic methylation and PAVs in four species (*B. oleracea*, *Solanum lycopersicum*, *Solanum tuberosum*, and *Zea mays*) with available PAV data (Figure S6; Table S17). For all genes (duplicated or not), teM genes were enriched amongst PAVs (FDR corrected $p < 0.001$) in all four species, while gbM was depleted (FDR corrected $p < 0.001$), except for *B. oleracea*. UnM genes were enriched in *S. lycopersicum* (FDR corrected $p < 0.001$) and depleted in the other three species. When limited to duplicate genes, the results were the same (Table S17). These results indicate that teM is associated with frequent gains or loss of genes and increased genetic variability, which over time can serve as an important source of genetic divergence and diversity.

**Genic methylation and divergence of paralog expression**

We explored how genic methylation relates to expression divergence between paralogs using gene expression atlases in *A. thaliana*, *Glycine max*, *Phaseolus vulgaris* and *Sorghum bicolor* (Klepikova et al. 2016; McCormick et al. 2018; Juexin Wang et al. 2019; O'Rourke et al. 2014). Most genes were expressed (Table S18) in at least one of these conditions (95.5%-99.4%), including the majority of teM genes (67.6%-98.5%). We then correlated expression between paralogs and plotted the distribution of these correlations based on the genic methylation of duplicate pairs (Figure S7). GbM-gbM pairs are the only duplicate pairs where the majority of paralogs are positively correlated in every species. TeM-containing duplicate pairs have distributions with two prominent peaks of ~0 (no correlation) and another peak near 1 (high

correlation). While low correlation, due to the role of teM in gene silencing, is expected for gbM-teM and unM-teM pairs, it was surprising that many of these pairs still maintained high-correlation. GbM-unM and unM-unM pairs had the most variable distribution patterns across species, ranging from mostly positive in *G. max* and *P. vulgaris,* to predominantly uncorrelated or bimodally distributed, like teM-containing pairs) in *S. bicolor* and *A. thaliana*. Notably, *G. max* shows the most distinct distributions, with the majority of paralogs positively correlated for all duplicate pairs. We suspect that the history of WGD is a major factor in all these correlations. For example, *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 (Schmutz et al. 2010)*. However, more extensive expression atlases and additional species will be needed to resolve these relationships.

We next examined the specificity of expression, that is how many conditions/tissues a gene is expressed, using τ (Tau) (Yanai et al. 2005). The value of τ ranges from '0' (broad expression) to '1' (narrow expression). Genes not expressed in any condition were removed as τ could not be calculated. GbM genes have the lowest τ, teM the highest, while unM genes have a breadth of intermediate values (Figure 5A, S8). 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 S9). As duplication types are enriched/depleted for different genic methylation, we examined the intersection of these factors on τ (Figure S9). Across all duplication types, τ was lowest in gbM and highest in teM. However, there were two unexpected results. Even though gbM genes generally have lower τ, local SGD gbM genes still showed a tendency to higher values compared to other duplicate types. Secondly, despite the fact that teM is associated with transcriptional silencing, WGD teM genes had a lower τ, and hence broader expression than SGD teM genes. This suggests that while genic methylation is an indicator of expression specificity, there are other factors related to the type of duplication that contributes to these patterns.

Duplicate pairs with the same genic methylation typically have lower absolute differences in τ than duplicate pairs that differed in genic methylation (Figure 5B, S10, Table S19). The greatest differences in τ were observed for gbM-teM pairs. However, unexpectedly gbM-unM genes often showed as great or greater difference in τ than unM-teM pairs. A possible explanation is that those paralogs most likely to gain or lose teM are already more narrowly expressed, leading to a smaller shift in τ. 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, S11). 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 gene-pairs 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 in either *G. max* or *P. vulgaris* gbM genes, although *G. max* gbM-unM had slightly higher τ (Figure S11). 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, indicating that these genes generally have broader expression than other teM-containing pairs. These data suggest that there may be some relationship between the parental expression of a gene and the expression of its duplicate copy and that certain genes may be more predisposed by their expression to certain switches in genic methylation.

## Genic methylation association with transposons and chromatin environment

Non-CG methylation is generally associated with TEs (Raju et al. 2019), so we next examined the association of paralogs with neighboring TEs. TE annotations were not available for all genomes, so we re-annotated the TEs of all 43 species using EDTA (Ou et al. 2019), ensuring a consistent methodology. We then examined each paralog for the presence of TEs within 1 kb of the gene body and tested for enrichment of TEs based on either genic methylation (Figure 6A, Table S20). TeM paralogs are enriched for TEs in the majority of species (36/43 enriched, 4/43 depleted), while unM paralogs were depleted for TEs in the majority of species (3/43 enriched, 33/43 depleted). No clear trend could be discerned for gbM paralogs, where similar numbers of species were enriched (15/43) or depleted (19/43) for TEs. This was not expected given the relationship between gbM and gene expression. We next examined TE enrichment based on the type of gene duplication and found a clear difference between WGD and SGDs (Figure 6B, Table S21). WGDs are depleted of TEs in the majority of species (2/43 enriched, 37/43 depleted), while all four types of SGDs, showed enrichment for TEs in the majority of species (Tandem: 30/43 enriched, 3/43 depleted; Proximal: 33/43 enriched, 2/43 depleted; Translocated: 21/43 enriched, 2/43 depleted; Dispersed: 27/43 enriched, 4/43 depleted). This relationship between TEs and SGDs may explain the enrichment of teM amongst SGDs in many species.

To further understand how TEs contribute to the dynamics of paralog genic methylation, we next examined the presence/absence of TEs for duplicate pairs differing in their genic methylation (Figure S12, Table S22). Our expectation was that for pairs in which one of the paralogs is teM, the teM pair would more frequently have a TE within 1 kb, while TEs would be absent for the non-teM paralog. However, this was not the case. Instead, for most species both paralogs in gbM-teM and unM-teM pairs are associated with a TE within 1 kb, while in *C. papaya* neither parlog was associated with a TE in the plurality of gbM-teM or unM-teM pairs. Only in *A. thaliana* did the plurality of both gbM-teM and unM-teM pairs show the expected pattern, while both *M. acuminata* and *E. guineensis* show the expected pattern for unM-teM pairs, but not gbM-teM pairs. Unexpectedly, for most species (37/43) both paralogs were associated with TEs for the plurality of gbM-unM pairs. While teM genes do show a greater association with TEs, these results suggest a more complex relationship than simple TE presence/absence in the switching of genic methylation states. This is especially true for species with larger genomes and greater TE-loads than *A. thaliana*.

Acquisition of teM could also be a factor of a gene's chromatin environment. Work in *G. max* has suggested that translocation of paralogs to TE-rich pericentromeric regions is a major source of teM genes (El Baidouri et al. 2018). To test if this pattern holds true across plant species, we examined genic methylation distribution throughout the genomes of each species. We used the number of genes, number of TEs, and number of nucleotides derived 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 6C, S13-15, Table S23). GbM, unM, and teM all showed positive correlations with the distribution of genes. The only exception was *A. thaliana*, where teM genes are negatively correlated with gene number (Pearson's r = ~-0.30, FDR corrected p-value < 0.001). This could be due in part to the genomic organization of *A. thaliana*, which has the smallest genome and the strongest negative correlation between total gene distribution and TEs (Pearson's r = ~-0.82, FDR corrected p-value < 0.001) and TE-base pairs (Pearson's r = ~-0.84, FDR corrected p-value < 0.001). In the majority of species, the distribution of gbM and unM genes was negatively correlated with both TE number (gbM: 8/43 positive, 34/43 negative; unM: 9/43 positive, 30/43 negative) and TE-base pairs

(gbM: 7/43 positive, 35/43 negative; unM: 8/43 positive, 35/43 negative), while teM genes were positively correlated with TEs (TE 28/43 positive, 10/43 negative) and TE-base pairs (26/43 positive, 15/43 negative). These results remained largely the same when restricted to duplicated genes (Table S23). Many of the species deviating from the expected pattern of teM distribution were in the Fabaceae (legumes) and Poaceae (grasses), prompting us to test these distributions for phylogenetic effects (Table S24). Both teM and gbM genes had significant phylogenetic signals for their correlations with the number of TEs and TE-base pairs. These significant phylogenetic signals for gbM remained even after the removal of *E. salsugineum*, *B. rapa*, and *B. oleracea* which have little to no gene-body methylation UnM and total genes showed a phylogenetic signal for their correlation with TE-base pairs, but not the number of TEs. As genome size correlates with genic non-CG methylation (Niederhuth et al. 2016; Takuno and Gaut 2012), we also tested for phylogenetic signals on genome size, but found none. These results indicate that there are lineage-specific differences in the distributions of genes, genic methylation, and TEs and that these differences are unlikely to be driven by genome size.

### Relationship of genic methylation frequency and sequence evolution within a population

Within a species DNA methylation can vary across the population (Becker and Weigel 2012). How this variation relates to paralog evolution is unknown. To address this, genes from 928 *A. thaliana* accessions in the 1001 Epigenomes Project (Kawakatsu et al. 2016) were binned based on the frequency of gbM/unM/teM in the population (0%, <25%, 25%-50%, 50%-75%, >75%). We then looked at the distribution of Ks and Ka/Ks to observe how sequence evolution related to genic methylation frequencies (Figure 7). Ks values decrease with increasing frequency of teM in the population, showing the biggest decrease when teM is above 50% in the population. However, even low frequency teM (<25%) genes tend to have lower Ks than genes with 0% teM. No obvious differences in Ks were observed across different frequencies of gbM or unM. Ka/Ks values increase with greater teM frequency, increase weakly with higher unM, and decrease with greater gbM frequency. Collectively these results indicate that the frequency of genic methylation states could differentially impact the evolution and divergence of paralogs within a population.

## Discussion

DNA methylation has long been proposed to play a role in the evolutionary fate of duplicate genes (Rodin and Riggs 2003; Keller and Yi 2014; Y. Wang et al. 2013; Jun Wang, Marowsky, and Fan 2014). However, this relationship has not been previously examined at either a phylogenetic or population level, leaving the generalizability of results from individual species unresolved. To address this issue, we examined DNA methylation and gene duplication across 43 angiosperms and a population of 928 *A. thaliana* accessions. Across species WGDs, local SGDs (tandem and proximal), and distal SGDs (translocated and dispersed) show general trends in their enrichment of genic methylation. However, there are notable exceptions, and we also identify phylogenetic effects on a number of these associations. For example, the depletion of gbM in WGDs of certain Brassiceae, which can be explained by the known depletion of gbM in these species (Bewick et al. 2016). Interestingly, a similar trend is observed in the Cucurbitaceae, despite no known depletion of gbM in these species, which needs further investigation.

To our knowledge, this is also the first study to examine the relationship between DNA methylation and paralog evolution across a population. Our results show differences in the sequence evolution of paralogs depending on the frequency of genic methylation states in the population. The frequency of genic methylation in the population may also provide clues as to when that genic methylation state was established. To achieve high frequency in a population, the simplest explanation is that a genic methylation state was established early following duplication, rather than being acquired individually multiple times. For example, in *A. thaliana* we argue that in the case of high-frequency teM, this genic methylation state was established early following gene duplication and maintained through subsequent diversification of *A. thaliana* accessions, while low frequency teM is more likely to have been acquired individually at various points.

Enrichment or depletion between types of duplication could occur either through shifts in DNA methylation states or biased amplification and retention of different classes of genes. Despite differing histories of WGD, we find that the majority of WGD pairs in all species shared the same genic methylation status. Similarly, studies of synthetic allopolyploids in both *Mimulus* and *Brassica* do not show extensive changes to DNA methylation in gene bodies, which would suggest that the parental state is largely maintained following allopolyploidy (Edger et al. 2017; Bird et al. 2021). As WGD results in duplication of the entire genome, it does not necessarily place duplicate genes in new sequence or chromatin contexts, which could explain the high degree of similarity of genic methylation between WGD pairs. In contrast to WGD, SGD is a continual process and places genes in potentially new sequence and chromatin environments. Extensive switching of genic methylation was previously observed for translocated duplicates in *G. max* (El Baidouri et al. 2018) and in our analyses translocated duplicates were more likely to differ in genic methylation than other duplicate types. Still, across all species, the majority of SGD pairs, including translocated duplicates, have the same genic methylation status. In some instances duplicate pairs could converge in genic methylation states by *trans*-acting mechanisms, such as RNA-directed DNA methylation (Raju et al. 2019), as observed in the case of the *A. thaliana PAI* gene family (Bender and Fink 1995). However, the simplest explanation is that contrary to previous assumptions, paralogs retain the same genic methylation state as the parental gene in most cases. This would then make biased amplification or retention the primary mechanism for the differences observed between duplication types.

Paralogs show distinct trends in the age of duplication and sequence evolution based on their genic methylation. GbM paralogs are more evolutionarily conserved in both sequence and expression, fitting with past observations of gbM genes (Takuno and Gaut 2013, 2012; Takuno, Ran, and Gaut 2016; Bewick et al. 2016), and would explain their retention following WGD. UnM genes are seemingly intermediate between gbM and teM in most aspects. UnM might be considered the 'default' state that spans from more gbM-like to more teM-like genes. For instance, in species that have lost gbM, the gbM ortholog is unM (Bewick et al. 2016). They are the largest of the three groups and broadly represented across both core angiosperm orthogroups and more lineage-specific orthogroups. Many transcription factors and kinases have tissue-specific expression, characteristic of unM, and are retained following WGD (Pophaly and Tellier 2015). At the same time, tandem and proximal duplication are often associated with environmental adaptation (Michael Freeling 2009). 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 will provide a constant source of new teM paralogs, leading to their enrichment. The combination of these processes would result in the observation that teM genes are evolutionarily younger. Transposons are the most likely explanation for teM paralogs, which show association with both local (<1 kb) TEs and TE-rich heterochromatic regions in most species. SGDs are also enriched for TEs in most species, which could lead to biased amplification of teM genes, further increasing their enrichment. Of course, we cannot ignore the possibility that some teM genes are misannotated transposons (Bennetzen et al. 2004; Schnable 2019). Annotation quality is a limitation of any genomics study and thorough reannotation of genomes is beyond the scope of this study. However, teM has been found in many known protein-coding gene families (Hua et al. 2013), including species-specific genes (Silveira et al. 2013). Furthermore, for translocated duplicates, the parental copy is a syntenic gene (Y. Wang, Li, and Paterson 2013), 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) (Rodin and Riggs 2003; Chang and Liao 2012; Adams et al. 2003). Alternatively, it is argued that silencing leads to pseudogenization and gene loss (Hua et al. 2013; El Baidouri et al. 2018). 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 suggest positive selection. Rapid functional divergence of SGDs was observed in grasses and many of these have characteristics similar to teM SGDs (Jiang and Assis 2019). Extensive expression divergence was observed in both teM and unM containing duplicate pairs, with only gbM-gbM pairs having mostly positive correlations of expressions in the species examined. DNA methylation in *cis*-regulatory regions (CREs) can also have an effect (Huang and Ecker 2018). However, genome-wide maps of CREs remain incomplete for most plant species, so we limited our analysis to coding regions. This systematic analysis reveals a number of general trends in the relationship between DNA methylation and gene duplication, as well as notable exceptions. For instance, we detect a phylogenetic effect on a number of associations, including distributions between TEs and genes. These exceptions could point to more interesting biology or be key to a deeper mechanistic understanding of this relationship.

# Methods:

## Genome and Methylome data

Genomes and gene annotations for 58 angiosperm species (Table S1) were used in this analysis (Garcia-Mas et al. 2012; Guo et al. 2013; Ming et al. 2008; Wu et al. 2018; Dohm et al. 2014; Parkin et al. 2014; Initiative and The International Brachypodium Initiative 2010; Amborella Genome Project 2013; Lamesch et al. 2012; C.-Y. Cheng et al. 2017; Bertioli et al. 2016; Hu et al. 2011; Sato et al. 2008; Paterson et al. 2012; Schmutz et al. 2010; Edger et al. 2019, 2018; Singh et al. 2013; Bartholomé et al. 2015; Q. Li et al. 2019; Slotte et al. 2013; D'Hont et al. 2012; R. Yang et al. 2013; Daccord et al. 2017; Bredeson et al. 2016; Hellsten et al. 2013; Tang et al. 2014; Kawahara et al. 2013; Lovell et al. 2018; Verde et al. 2017; Tuskan et al. 2006;

Schmutz et al. 2014; Xue et al. 2018; McCormick et al. 2018; Bennetzen et al. 2012; Hosmani et al., n.d.; Sharma et al. 2013; Mamidi et al., n.d.; Motamayor et al. 2013; Jiao et al. 2017; Hibrand Saint-Oyant et al. 2018; VanBuren et al. 2018; Liu et al. 2014; Colle et al. 2019; VanBuren et al. 2015; Harkess et al. 2017; Hulse-Kemp et al. 2018; S. Xu et al. 2017; Bombarely et al. 2016; Ming et al. 2013; W. Wang et al. 2014; Jaillon et al. 2007; Valliyodan et al. 2019; Filiault et al. 2018; Lovell et al. 2021; Barchi et al. 2019). This includes 43 species (Table S1) with whole-genome bisulfite sequencing (WGBS) data (Amborella Genome Project 2013; Seymour et al. 2014; Picard and Gehring 2017; Bertioli et al. 2016; Niederhuth et al. 2016; Bewick et al. 2016; Lü et al. 2018; Ong-Abdullah et al. 2015; J. Cheng et al. 2018; Kim et al. 2015; Song et al. 2017; Daccord et al. 2017; Secco et al. 2015; Dong et al. 2017; Y. Yang et al. 2019; L. Wang et al. 2018; Turco et al. 2017; Noshay et al. 2019) and an additional 11 species which were included as outgroups for orthogroup analysis and *MCScanX-Transposed*. Genome data was individually downloaded from multiple databases (Table S1) (Goodstein et al. 2012; Portwood et al. 2019; Howe et al. 2020; Jung et al. 2019; Dash et al. 2016; Zheng et al. 2019; Lyons and Freeling 2008; Fernandez-Pozo et al. 2015). Protein fastas, CDS fastas, and annotation files were filtered to retain only the primary transcript.

**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 (Schultz et al. 2015). 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 that context across coding regions of all species (Niederhuth et al. 2016). Each gene was tested for enrichment of CG, CHG, CHH, or 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 (Benjamini and Hochberg 1995). Genes were classified as gene-body methylated (gbM), TE-like methylated (teM), and unmethylated (unM) based on their genic methylation as previously described (Niederhuth et al. 2016; Takuno and Gaut 2012). Genes enriched for CG methylation with ≥ 10 CG sites and non-significant CHG or CHH methylation were classified as gbM. Genes enriched for CHG, CHH, or non-CG and ≥ 10 sites in that context were classified as teM. Genes with ≤ 1 methylated site in any context or a weighted methylation (Schultz, Schmitz, and Ecker 2012) for all contexts (CG, CHG, or CHH) ≤ 2% were classified as unM. Genes lacking DNA methylation data were considered 'missing' and those with intermediate DNA methylation levels not fitting the above criteria as 'unclassified'.

**Orthogroup analyses**

Protein sequences from 58 angiosperm species (See Supporting Information, Table S1) were classified into orthogroups using Orthofinder v2.5.2 (Emms and Kelly 2019, 2015), with the options '-M dendroblast -S diamond_ultra_sens, -I 1.3'. 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%) (Z. Li et al. 2016). Following Li et al., we further classified core angiosperm orthogroups as single-copy if represented by a single gene in ≥ 70% species. Non-core orthogroups were classified as "cross-family" (present in ≥ 2 species from

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 (Buchfink, Xie, and Huson 2015) was used to perform a blastp against itself and *A. trichopoda*, retaining hits with e-value < 1e-5. For *A. trichopoda*, *A. thaliana* was the outgroup. Blastp results were filtered to remove hits from differing orthogroups. Duplicate genes were classified by *DupGen_finder-unique (Qiao et al. 2019)*, requiring ≥ 5 genes for collinearity and ≤ 10 intervening genes to classify as 'proximal' duplicates. *MCScanX-transposed (Y. Wang, Li, and Paterson 2013)* was used to detect translocated duplicates occurring within different epochs since species divergence (Table S13). Enrichment between duplication type and genic methylation was determined by a two-sided Fisher's exact test (Fisher 1934) with FDR-correction by BH and plotted using *heatmap.2* in *gplots* ("Various R Programming Tools for Plotting Data [R Package Gplots Version 3.1.1]" 2020). The phylogenetic tree in Figure 1 was created using 'V.PhyloMaker' (Jin and Qian 2019) and 'phytools' (Revell 2012).

**Sequence evolution**

The calculate_Ka_Ks_pipeline.pl (Qiao et al. 2019) was used to determine nonsynonymous (Ka) and synonymous substitutions (Ks) for duplicate pairs. Protein sequences are aligned by MAFFT (v7.402) (Katoh and Standley 2013), converted to a codon alignment with PAL2NAL (Suyama, Torrents, and Bork 2006), and KaKs_Calculator 2.0 used to calculate Ka, Ks, and Ka/Ks with the γ-MYN method (D. Wang et al. 2010; Qiao et al. 2019). 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 (Massey 1951) with FDR-correction by BH. PAV variants were downloaded for *B. oleracea (Golicz et al. 2016)*, *S. lycopersicum (Gao et al. 2019)*, *S. tuberosum (Hardigan et al. 2016)*, and *Z. mays (Hirsch et al. 2014)*. 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 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* are from published data (Klepikova et al. 2016; McCormick et al. 2018; Juexin Wang et al. 2019; O'Rourke et al. 2014). Raw data for *A. thaliana* was downloaded from NCBI SRA (PRJNA314076 and PRJNA324514), mapped using STAR (Dobin et al. 2013), and normalized by DESeq2 (Love, Huber, and Anders 2014). For *G. max*, *P. vulgaris*, and *S.bicolor*, normalized data was downloaded from Phytozome (Goodstein et al. 2012). Pearson correlation coefficients were calculated for each duplicate pair and the tissue-specificity index (τ) (Yanai et al. 2005) calculated for each gene.

**Transposons and genomic distribution**

TEs were called *de novo* for all species using the Extensive *de-novo* TE Annotator pipeline (Ou et al. 2019). We calculated the total number of genes, genes belonging to each of the genic

methylation classes, the number of TEs, and number of TE base pairs in 100 kb sliding windows with 50 kb steps. Pearson correlation coefficients were calculated using the '*rcorr*' function in '*corrplot*' (Wei and Viliam n.d.). Genome sizes were obtained from the Plant DNA C-value database (Pellicer and Leitch 2020) (release 7.1).

### Phylogenetic signal

Phylogenetic signal was tested using Pagel's lambda (Pagel 1999) using *phylosig* in *phytools (Revell 2012)*. The input phylogenetic tree (Dataset S1) was generated with orthofinder (Emms and Kelly 2015). For enrichment/depletion of genic methylation in different duplication types, statistically significant depletion was coded as -1, enrichment 1, and non-significant results 0 before testing for phylogenetic signal.

### Arabidopsis diversity

WGBS data for 928 accessions from the *Arabidopsis thaliana* 1001 Epigenomes Project (Kawakatsu et al. 2016), previously aligned by methylpy, was downloaded from the Gene Expression Omnibus (GEO Accession GSE43857).Genes were classified as before and the frequency of each genic methylation class for each gene in the population calculated.

### Data availability and research reproducibility

Data used are listed in the Supporting Information and Table-S1. Formatted genomes and annotations for replication are available at DataDryad XX. Code is available at: https://github.com/niederhuth/DNA-methylation-signatures-of-duplicate-gene-evolution-in-angiosperms.

**Acknowledgements:** We thank Dr. Patrick Edger for the unpublished *C. violaceae* genome and Dr. Leslie Kollar for reviewing the 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 (NSF grant DBI-1757043).

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

## Figure Legends:

**Figure 1: 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 2: 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 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). Box plots (C and D) show the distribution of synonymous substitutions (Ks) for each of the duplicate-pair genic methylation classifications in *Brachypodium distachyon* and *Phaseolus vulgaris* respectively.

**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 *Brachypodium distachyon* and *Phaseolus 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: Local and genome-wide transposon and chromatin environment associations of duplicate genes.** A) Enrichment and depletion of transposable elements (TEs) with gbM, teM, and unM paralogs and different types of duplication in each species. TEs within 1 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. B) Genomic features such as number of genes and number of TEs were calculated in 100kb sliding windows with a 50kb step size. Increasing shades of red indicate positive correlation, while increasing shades of blue represent negative correlations. Unless indicated, all associations are statistically significant at a FDR-corrected p-value < 0.05. 'NS' indicates no statistical significance.

**Figure 7: Genic methylation frequency in a population is associated with age of duplication and sequence evolution.** A) Density plots showing the Ks distribution of genes at different frequencies of gbM, unM, and teM (0%, <25%, 25%-50%, 50%-75%, >75%) in the population. 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 orthogroups and genic methylation classes. A) 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). B) Distribution of genes classified as gbM, unM, teM, unclassified, and 'missing methylation data' across different orthogroup categories. If a plant family is only represented by a single species, family and species specific genes are group together as species specific.

**Figure S3:** 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 S4:** 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 S5:** 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 S6:** 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 S7:** 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 S8**: Tau specificity of gbM, teM, and unM genes in *G. max*, *P. vulgaris*, and *S. bicolor*.

**Figure S9:** 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 S10:** 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 S11:** 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 S12:** Proportion of duplicate pairs with or without a transposable element (TE) in the immediate vicinity of the genes (gene body, 1kb up and 1 kb downstream). Presence of a TE is indicated by '1' and absence with a '0'. For example, a duplicate pair, where paralog 1 has a TE present in the immediate vicinity and paralog 2 doesnt have any TE's is indicated by '1-0'

**Figure S13:** Pearson correlations between genic methylation classes (gbM, unM, teM) and genomic features (number of genes, and TE base pairs). Number of genes, TEs and TE base pairs 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 base pairs, 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*.

## 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 orange, depleted associations are in green. FDR corrected 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 corrected 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. Fisher exact test odds ratios of enriched associations are colored orange, depleted associations are in green. Blue indicates distribution is significantly different at an FDR corrected p-value < 0.05.

**Table S9:** Number of genes with similar or divergent methylation profiles between parental and translocated duplicates.

**Table S10:** Duplicate pair classifications ranked based on median Ks values for single gene duplicates (SGD) and whole-genome duplicates (WGD).

**Table S11:** 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 corrected p-value < 0.05.

**Table S12:** Outgroup species used for each epoch as part of MCscanX-transposed.

**Table S13:** 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 orange, depleted associations are in green. Blue indicates distribution is significantly different at an FDR corrected p-value < 0.05.

**Table S14:** Duplicate pair classifications ranked based on median Ka/Ks values for single gene duplicates (SGD) and whole-genome duplicates (WGD).

**Table S15:** 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 corrected p-value < 0.05.

**Table S16:** Enrichment and depletion of different classifications of duplicate gene pairs with Ka/Ks ratio > 1.0. Odds ratios of enriched associations are colored orange, depleted associations are in green. Blue indicates distribution is significantly different at a 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 orange, depleted associations are in green. Blue indicates distribution is significantly different at a FDR corrected p-value < 0.05.

**Table S18:** Number of genes with no detectable expression in any tissue/treatments in the gene expression atlases.

**Table S19**: Differences in the distribution of Tau - absolute difference for duplicate gene pairs with or without methylation divergence. ANOVA tests followed by Tukey's HSD were computed to find significant differences between duplicate pair classifications (marked in blue).

**Table S20:** Enrichment and depletion of transposable elements (TEs) with gbM, teM, and unM paralogs in each species. TEs within 1 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 (orange). Associations are considered significant at a FDR corrected p-value 0.05 and shown 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 corrected p-value < 0.05 and are highlighted in yellow.

**Table S21:** Enrichment and depletion of transposable elements (TEs) with different types of duplication in each species. TEs within 1 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 (orange). Associations are considered significant at a FDR corrected p-value < 0.05 and shown 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 corrected p-value < 0.05 and are highlighted in yellow.

**Table S22:** Presence/Absence of TEs in the gene body and 1 kb upstream and 1 kb downstream for duplicate paralogs differing in their genic methylation.

**Table S23:** Correlations between genic methylation classes (gbM, teM, unM) and genomic features (number of genes, TEs, and TE base pairs) in 100kb sliding windows with a 50kb step size. Positive correlations are marked in orange, negative correlations in green. Blue indicates distribution is significantly different at a FDR corrected p-value < 0.05.

**Table S24:** Pagel's lambda test for phylogenetic signal of correlations in Table S23. 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 corrected p < 0.05.

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

## References:

Adams, K. L., R. Cronn, R. Percifield, and J. F. Wendel. 2003. "Genes Duplicated by Polyploidy Show Unequal Contributions to the Transcriptome and Organ-Specific Reciprocal Silencing." *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.0630618100.

Amborella Genome Project. 2013. "The Amborella Genome and the Evolution of Flowering Plants." *Science* 342 (6165): 1241089.

Barchi, Lorenzo, Marco Pietrella, Luca Venturini, Andrea Minio, Laura Toppino, Alberto Acquadro, Giuseppe Andolfo, et al. 2019. "A Chromosome-Anchored Eggplant Genome Sequence Reveals Key Events in Solanaceae Evolution." *Scientific Reports* 9 (1): 11769.

Bartholomé, Jérôme, Eric Mandrou, André Mabiala, Jerry Jenkins, Ibouniyamine Nabihoudine, Christophe Klopp, Jeremy Schmutz, Christophe Plomion, and Jean-Marc Gion. 2015. "High-Resolution Genetic Maps of Eucalyptus Improve Eucalyptus Grandis Genome Assembly." *The New Phytologist* 206 (4): 1283–96.

Becker, Claude, and Detlef Weigel. 2012. "Epigenetic Variation: Origin and Transgenerational Inheritance." *Current Opinion in Plant Biology* 15 (5): 562–67.

Bender, J., and G. R. Fink. 1995. "Epigenetic Control of an Endogenous Gene Family Is Revealed by a Novel Blue Fluorescent Mutant of Arabidopsis." *Cell* 83 (5): 725–34.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical

and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society: Series B (Methodological)*. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Bennetzen, Jeffrey L., Craig Coleman, Renyi Liu, Jianxin Ma, and Wusirika Ramakrishna. 2004. "Consistent over-Estimation of Gene Number in Complex Plant Genomes." *Current Opinion in Plant Biology* 7 (6): 732–36.

Bennetzen, Jeffrey L., Jeremy Schmutz, Hao Wang, Ryan Percifield, Jennifer Hawkins, Ana C. Pontaroli, Matt Estep, et al. 2012. "Reference Genome Sequence of the Model Plant Setaria." *Nature Biotechnology* 30 (6): 555–61.

Bertioli, David John, Steven B. Cannon, Lutz Froenicke, Guodong Huang, Andrew D. Farmer, Ethalinda K. S. Cannon, Xin Liu, et al. 2016. "The Genome Sequences of Arachis Duranensis and Arachis Ipaensis, the Diploid Ancestors of Cultivated Peanut." *Nature Genetics* 48 (4): 438–46.

Bewick, Adam J., Lexiang Ji, Chad E. Niederhuth, Eva-Maria Willing, Brigitte T. Hofmeister, Xiuling Shi, Li Wang, et al. 2016. "On the Origin and Evolutionary Consequences of Gene Body DNA Methylation." *Proceedings of the National Academy of Sciences of the United States of America* 113 (32): 9111–16.

Bird, Kevin A., Chad E. Niederhuth, Shujun Ou, Malia Gehan, J. Chris Pires, Zhiyong Xiong, Robert VanBuren, and Patrick P. Edger. 2021. "Replaying the Evolutionary Tape to Investigate Subgenome Dominance in Allopolyploid Brassica Napus." *The New Phytologist* 230 (1): 354–71.

Bombarely, Aureliano, Michel Moser, Avichai Amrad, Manzhu Bao, Laure Bapaume, Cornelius S. Barry, Mattijs Bliek, et al. 2016. "Insight into the Evolution of the Solanaceae from the Parental Genomes of Petunia Hybrida." *Nature Plants* 2 (6): 16074.

Bredeson, Jessen V., Jessica B. Lyons, Simon E. Prochnik, G. Albert Wu, Cindy M. Ha, Eric Edsinger-Gonzales, Jane Grimwood, et al. 2016. "Sequencing Wild and Cultivated Cassava and Related Species Reveals Extensive Interspecific Hybridization and Genetic Diversity." *Nature Biotechnology* 34 (5): 562–70.

Bridges, Calvin B. 1935. "SALIVARY CHROMOSOME MAPS." *Journal of Heredity*. https://doi.org/10.1093/oxfordjournals.jhered.a104022.

Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. 2015. "Fast and Sensitive Protein Alignment Using DIAMOND." *Nature Methods*. https://doi.org/10.1038/nmeth.3176.

Chang, Andrew Ying-Fei, and Ben-Yang Liao. 2012. "DNA Methylation Rebalances Gene Dosage after Mammalian Gene Duplications." *Molecular Biology and Evolution* 29 (1): 133–44.

Cheng, Chia-Yi, Vivek Krishnakumar, Agnes P. Chan, Françoise Thibaud-Nissen, Seth Schobel, and Christopher D. Town. 2017. "Araport11: A Complete Reannotation of the Arabidopsis Thaliana Reference Genome." *The Plant Journal: For Cell and Molecular Biology* 89 (4): 789–804.

Cheng, Feng, Jian Wu, Xu Cai, Jianli Liang, Michael Freeling, and Xiaowu Wang. 2018. "Gene Retention, Fractionation and Subgenome Differences in Polyploid Plants." *Nature Plants* 4 (5): 258–68.

Cheng, Jingfei, Qingfeng Niu, Bo Zhang, Kunsong Chen, Ruihua Yang, Jian-Kang Zhu, Yijing Zhang, and Zhaobo Lang. 2018. "Downregulation of RdDM during Strawberry Fruit Ripening." *Genome Biology* 19 (1): 212.

Colle, Marivi, Courtney P. Leisner, Ching Man Wai, Shujun Ou, Kevin A. Bird, Jie Wang, Jennifer H. Wisecaver, et al. 2019. "Haplotype-Phased Genome and Evolution of

Phytonutrient Pathways of Tetraploid Blueberry." *GigaScience* 8 (3). https://doi.org/10.1093/gigascience/giz012.

Consortium, The Potato Genome Sequencing, and The Potato Genome Sequencing Consortium. 2011. "Genome Sequence and Analysis of the Tuber Crop Potato." *Nature*. https://doi.org/10.1038/nature10158.

Cusack, Brian P., and Kenneth H. Wolfe. 2007. "Not Born Equal: Increased Rate Asymmetry in Relocated and Retrotransposed Rodent Gene Duplicates." *Molecular Biology and Evolution* 24 (3): 679–86.

Daccord, Nicolas, Jean-Marc Celton, Gareth Linsmith, Claude Becker, Nathalie Choisne, Elio Schijlen, Henri van de Geest, et al. 2017. "High-Quality de Novo Assembly of the Apple Genome and Methylome Dynamics of Early Fruit Development." *Nature Genetics* 49 (7): 1099–1106.

Dash, Sudhansu, Ethalinda K. S. Cannon, Scott R. Kalberer, Andrew D. Farmer, and Steven B. Cannon. 2016. "PeanutBase and Other Bioinformatic Resources for Peanut." *Peanuts*. https://doi.org/10.1016/b978-1-63067-038-2.00008-3.

De Smet, Riet, Keith L. Adams, Klaas Vandepoele, Marc C. E. Van Montagu, Steven Maere, and Yves Van de Peer. 2013. "Convergent Gene Loss Following Gene and Genome Duplications Creates Single-Copy Families in Flowering Plants." *Proceedings of the National Academy of Sciences of the United States of America* 110 (8): 2898–2903.

D'Hont, Angélique, France Denoeud, Jean-Marc Aury, Franc-Christophe Baurens, Françoise Carreel, Olivier Garsmeur, Benjamin Noel, et al. 2012. "The Banana (Musa Acuminata) Genome and the Evolution of Monocotyledonous Plants." *Nature* 488 (7410): 213–17.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics* 29 (1): 15–21.

Dohm, Juliane C., Cornelia Lange, Daniela Holtgräwe, Thomas Rosleff Sörensen, Dietrich Borchardt, Britta Schulz, Hans Lehrach, Bernd Weisshaar, and Heinz Himmelbauer. 2012. "Palaeohexaploid Ancestry for Caryophyllales Inferred from Extensive Gene-Based Physical and Genetic Mapping of the Sugar Beet Genome (Beta Vulgaris)." *The Plant Journal: For Cell and Molecular Biology* 70 (3): 528–40.

Dohm, Juliane C., André E. Minoche, Daniela Holtgräwe, Salvador Capella-Gutiérrez, Falk Zakrzewski, Hakim Tafer, Oliver Rupp, et al. 2014. "The Genome of the Recently Domesticated Crop Plant Sugar Beet (Beta Vulgaris)." *Nature* 505 (7484): 546–49.

Dong, Pengfei, Xiaoyu Tu, Po-Yu Chu, Peitao Lü, Ning Zhu, Donald Grierson, Baijuan Du, Pinghua Li, and Silin Zhong. 2017. "3D Chromatin Architecture of Large Plant Genomes Determined by Local A/B Compartments." *Molecular Plant*. https://doi.org/10.1016/j.molp.2017.11.005.

Edger, Patrick P., Thomas J. Poorten, Robert VanBuren, Michael A. Hardigan, Marivi Colle, Michael R. McKain, Ronald D. Smith, et al. 2019. "Origin and Evolution of the Octoploid Strawberry Genome." *Nature Genetics* 51 (3): 541–47.

Edger, Patrick P., Ronald Smith, Michael R. McKain, Arielle M. Cooley, Mario Vallejo-Marin, Yaowu Yuan, Adam J. Bewick, et al. 2017. "Subgenome Dominance in an Interspecific Hybrid, Synthetic Allopolyploid, and a 140-Year-Old Naturally Established Neo-Allopolyploid Monkeyflower." *The Plant Cell*. https://doi.org/10.1105/tpc.17.00010.

Edger, Patrick P., Robert VanBuren, Marivi Colle, Thomas J. Poorten, Ching Man Wai, Chad E. Niederhuth, Elizabeth I. Alger, et al. 2018. "Single-Molecule Sequencing and Optical

Mapping Yields an Improved Genome of Woodland Strawberry (Fragaria Vesca) with Chromosome-Scale Contiguity." *GigaScience* 7 (2): 1–7.

El Baidouri, Moaine, Kyung Do Kim, Brian Abernathy, Ying-Hui Li, Li-Juan Qiu, and Scott A. Jackson. 2018. "Genic C-Methylation in Soybean Is Associated with Gene Paralogs Relocated to Transposable Element-Rich Pericentromeres." *Molecular Plant* 11 (3): 485–95.

Emms, David M., and Steven Kelly. 2015. "OrthoFinder: Solving Fundamental Biases in Whole Genome Comparisons Dramatically Improves Orthogroup Inference Accuracy." *Genome Biology* 16 (August): 157.

———. 2019. "OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics." *Genome Biology* 20 (1): 238.

Feng, Suhua, Shawn J. Cokus, Xiaoyu Zhang, Pao-Yang Chen, Magnolia Bostick, Mary G. Goll, Jonathan Hetzel, et al. 2010. "Conservation and Divergence of Methylation Patterning in Plants and Animals." *Proceedings of the National Academy of Sciences of the United States of America* 107 (19): 8689–94.

Fernandez-Pozo, Noe, Naama Menda, Jeremy D. Edwards, Surya Saha, Isaak Y. Tecle, Susan R. Strickler, Aureliano Bombarely, et al. 2015. "The Sol Genomics Network (SGN)--from Genotype to Phenotype to Breeding." *Nucleic Acids Research* 43 (Database issue): D1036–41.

Filiault, Danièle L., Evangeline S. Ballerini, Terezie Mandáková, Gökçe Aköz, Nathan J. Derieg, Jeremy Schmutz, Jerry Jenkins, et al. 2018. "The Genome Provides Insight into Adaptive Radiation and Reveals an Extraordinarily Polymorphic Chromosome with a Unique History." *eLife* 7 (October). https://doi.org/10.7554/eLife.36426.

Fisher, Sir Ronald Aylmer. 1934. *Statistical Methods for Research Workers.*

Flagel, Lex E., and Jonathan F. Wendel. 2009. "Gene Duplication and Evolutionary Novelty in Plants." *New Phytologist*. https://doi.org/10.1111/j.1469-8137.2009.02923.x.

Freeling, Michael. 2009. "Bias in Plant Gene Content Following Different Sorts of Duplication: Tandem, Whole-Genome, Segmental, or by Transposition." *Annual Review of Plant Biology*. https://doi.org/10.1146/annurev.arplant.043008.092122.

Freeling, M., E. Lyons, B. Pedersen, M. Alam, R. Ming, and D. Lisch. 2008. "Many or Most Genes in Arabidopsis Transposed after the Origin of the Order Brassicales." *Genome Research*. https://doi.org/10.1101/gr.081026.108.

Ganko, Eric W., Blake C. Meyers, and Todd J. Vision. 2007. "Divergence in Expression between Duplicated Genes in Arabidopsis." *Molecular Biology and Evolution* 24 (10): 2298–2309.

Gao, Lei, Itay Gonda, Honghe Sun, Qiyue Ma, Kan Bao, Denise M. Tieman, Elizabeth A. Burzynski-Chang, et al. 2019. "The Tomato Pan-Genome Uncovers New Genes and a Rare Allele Regulating Fruit Flavor." *Nature Genetics* 51 (6): 1044–51.

Garcia-Mas, Jordi, Andrej Benjak, Walter Sanseverino, Michael Bourgeois, Gisela Mir, Víctor M. González, Elizabeth Hénaff, et al. 2012. "The Genome of Melon (Cucumis Melo L.)." *Proceedings of the National Academy of Sciences of the United States of America* 109 (29): 11872–77.

Golicz, Agnieszka A., Philipp E. Bayer, Guy C. Barker, Patrick P. Edger, Hyeran Kim, Paula A. Martinez, Chon Kit Kenneth Chan, et al. 2016. "The Pangenome of an Agronomically Important Crop Plant Brassica Oleracea." *Nature Communications* 7 (November): 13390.

Goodstein, David M., Shengqiang Shu, Russell Howson, Rochak Neupane, Richard D. Hayes, Joni Fazo, Therese Mitros, et al. 2012. "Phytozome: A Comparative Platform for Green

Plant Genomics." *Nucleic Acids Research* 40 (Database issue): D1178–86.
Guo, Shaogui, Jianguo Zhang, Honghe Sun, Jerome Salse, William J. Lucas, Haiying Zhang, Yi Zheng, et al. 2013. "The Draft Genome of Watermelon (Citrullus Lanatus) and Resequencing of 20 Diverse Accessions." *Nature Genetics* 45 (1): 51–58.
Hardigan, Michael A., Emily Crisovan, John P. Hamilton, Jeongwoon Kim, Parker Laimbeer, Courtney P. Leisner, Norma C. Manrique-Carpintero, et al. 2016. "Genome Reduction Uncovers a Large Dispensable Genome and Adaptive Role for Copy Number Variation in Asexually Propagated Solanum Tuberosum." *The Plant Cell* 28 (2): 388–405.
Harkess, Alex, Jinsong Zhou, Chunyan Xu, John E. Bowers, Ron Van der Hulst, Saravanaraj Ayyampalayam, Francesco Mercati, et al. 2017. "The Asparagus Genome Sheds Light on the Origin and Evolution of a Young Y Chromosome." *Nature Communications*. https://doi.org/10.1038/s41467-017-01064-8.
Hellsten, Uffe, Kevin M. Wright, Jerry Jenkins, Shengqiang Shu, Yaowu Yuan, Susan R. Wessler, Jeremy Schmutz, John H. Willis, and Daniel S. Rokhsar. 2013. "Fine-Scale Variation in Meiotic Recombination in Mimulus Inferred from Population Shotgun Sequencing." *Proceedings of the National Academy of Sciences of the United States of America* 110 (48): 19478–82.
Hibrand Saint-Oyant, L., T. Ruttink, L. Hamama, I. Kirov, D. Lakhwani, N. N. Zhou, P. M. Bourke, et al. 2018. "A High-Quality Genome Sequence of Rosa Chinensis to Elucidate Ornamental Traits." *Nature Plants* 4 (7): 473–84.
Hirsch, Candice N., Jillian M. Foerster, James M. Johnson, Rajandeep S. Sekhon, German Muttoni, Brieanne Vaillancourt, Francisco Peñagaricano, et al. 2014. "Insights into the Maize Pan-Genome and Pan-Transcriptome." *The Plant Cell* 26 (1): 121–35.
Hosmani, Prashant S., Mirella Flores-Gonzalez, Henri van de Geest, Florian Maumus, Linda V. Bakker, Elio Schijlen, Jan van Haarst, et al. n.d. "An Improved de Novo Assembly and Annotation of the Tomato Reference Genome Using Single-Molecule Sequencing, Hi-C Proximity Ligation and Optical Maps." https://doi.org/10.1101/767764.
Howe, Kevin L., Bruno Contreras-Moreira, Nishadi De Silva, Gareth Maslen, Wasiu Akanni, James Allen, Jorge Alvarez-Jarreta, et al. 2020. "Ensembl Genomes 2020-Enabling Non-Vertebrate Genomic Research." *Nucleic Acids Research* 48 (D1): D689–95.
Huang, Shao-Shan C., and Joseph R. Ecker. 2018. "Piecing Together Cis-Regulatory Networks: Insights from Epigenomics Studies in Plants." *Wiley Interdisciplinary Reviews. Systems Biology and Medicine* 10 (3): e1411.
Hua, Zhihua, John E. Pool, Robert J. Schmitz, Matthew D. Schultz, Shin-Han Shiu, Joseph R. Ecker, and Richard D. Vierstra. 2013. "Epigenomic Programming Contributes to the Genomic Drift Evolution of the F-Box Protein Superfamily in Arabidopsis." *Proceedings of the National Academy of Sciences of the United States of America* 110 (42): 16927–32.
Hulse-Kemp, Amanda M., Shamoni Maheshwari, Kevin Stoffel, Theresa A. Hill, David Jaffe, Stephen R. Williams, Neil Weisenfeld, et al. 2018. "Reference Quality Assembly of the 3.5-Gb Genome of Capsicum Annuum from a Single Linked-Read Library." *Horticulture Research*. https://doi.org/10.1038/s41438-017-0011-0.
Hu, Tina T., Pedro Pattyn, Erica G. Bakker, Jun Cao, Jan-Fang Cheng, Richard M. Clark, Noah Fahlgren, et al. 2011. "The Arabidopsis Lyrata Genome Sequence and the Basis of Rapid Genome Size Change." *Nature Genetics* 43 (5): 476–81.
Initiative, The International Brachypodium, and The International Brachypodium Initiative. 2010. "Genome Sequencing and Analysis of the Model Grass Brachypodium Distachyon."

*Nature*. https://doi.org/10.1038/nature08747.

Jaillon, Olivier, Jean-Marc Aury, Benjamin Noel, Alberto Policriti, Christian Clepet, Alberto Casagrande, Nathalie Choisne, et al. 2007. "The Grapevine Genome Sequence Suggests Ancestral Hexaploidization in Major Angiosperm Phyla." *Nature* 449 (7161): 463–67.

Jiang, Xueyuan, and Raquel Assis. 2019. "Rapid Functional Divergence after Small-Scale Gene Duplication in Grasses." *BMC Evolutionary Biology* 19 (1): 97.

Jiao, Yinping, Paul Peluso, Jinghua Shi, Tiffany Liang, Michelle C. Stitzer, Bo Wang, Michael S. Campbell, et al. 2017. "Improved Maize Reference Genome with Single-Molecule Technologies." *Nature* 546 (7659): 524–27.

Jin, Yi, and Hong Qian. 2019. "V.PhyloMaker: An R Package That Can Generate Very Large Phylogenies for Vascular Plants." *Ecography*. https://doi.org/10.1111/ecog.04434.

Jung, Sook, Taein Lee, Chun-Huai Cheng, Katheryn Buble, Ping Zheng, Jing Yu, Jodi Humann, et al. 2019. "15 Years of GDR: New Data and Functionality in the Genome Database for Rosaceae." *Nucleic Acids Research* 47 (D1): D1137–45.

Katoh, Kazutaka, and Daron M. Standley. 2013. "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability." *Molecular Biology and Evolution* 30 (4): 772–80.

Kawahara, Yoshihiro, Melissa de la Bastide, John P. Hamilton, Hiroyuki Kanamori, W. Richard McCombie, Shu Ouyang, David C. Schwartz, et al. 2013. "Improvement of the Oryza Sativa Nipponbare Reference Genome Using next Generation Sequence and Optical Map Data." *Rice* 6 (1): 4.

Kawakatsu, Taiji, Shao-Shan Carol Huang, Florian Jupe, Eriko Sasaki, Robert J. Schmitz, Mark A. Urich, Rosa Castanon, et al. 2016. "Epigenomic Diversity in a Global Collection of Arabidopsis Thaliana Accessions." *Cell* 166 (2): 492–505.

Keller, T. E., and S. V. Yi. 2014. "DNA Methylation and Evolution of Duplicate Genes." *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1321420111.

Kim, Kyung Do, Moaine El Baidouri, Brian Abernathy, Aiko Iwata-Otsubo, Carolina Chavarro, Michael Gonzales, Marc Libault, Jane Grimwood, and Scott A. Jackson. 2015. "A Comparative Epigenomic Analysis of Polyploidy-Derived Genes in Soybean and Common Bean." *Plant Physiology* 168 (4): 1433–47.

Klepikova, Anna V., Artem S. Kasianov, Evgeny S. Gerasimov, Maria D. Logacheva, and Aleksey A. Penin. 2016. "A High Resolution Map of the Arabidopsis Thaliana Developmental Transcriptome Based on RNA-Seq Profiling." *The Plant Journal: For Cell and Molecular Biology* 88 (6): 1058–70.

Lamesch, Philippe, Tanya Z. Berardini, Donghui Li, David Swarbreck, Christopher Wilks, Rajkumar Sasidharan, Robert Muller, et al. 2012. "The Arabidopsis Information Resource (TAIR): Improved Gene Annotation and New Tools." *Nucleic Acids Research* 40 (Database issue): D1202–10.

Li, Qing, Hongbo Li, Wu Huang, Yuanchao Xu, Qian Zhou, Shenhao Wang, Jue Ruan, Sanwen Huang, and Zhonghua Zhang. 2019. "A Chromosome-Scale Genome Assembly of Cucumber (Cucumis Sativus L.)." *GigaScience* 8 (6). https://doi.org/10.1093/gigascience/giz072.

Liu, Meng-Jun, Jin Zhao, Qing-Le Cai, Guo-Cheng Liu, Jiu-Rui Wang, Zhi-Hui Zhao, Ping Liu, et al. 2014. "The Complex Jujube Genome Provides Insights into Fruit Tree Biology." *Nature Communications* 5 (October): 5315.

Li, Zhen, Jonas Defoort, Setareh Tasdighian, Steven Maere, Yves Van de Peer, and Riet De Smet. 2016. “Gene Duplicability of Core Genes Is Highly Consistent across All Angiosperms.” *The Plant Cell* 28 (2): 326–44.

Lovell, John T., Jerry Jenkins, David B. Lowry, Sujan Mamidi, Avinash Sreedasyam, Xiaoyu Weng, Kerrie Barry, et al. 2018. “The Genomic Landscape of Molecular Responses to Natural Drought Stress in Panicum Hallii.” *Nature Communications* 9 (1): 5213.

Lovell, John T., Alice H. MacQueen, Sujan Mamidi, Jason Bonnette, Jerry Jenkins, Joseph D. Napier, Avinash Sreedasyam, et al. 2021. “Genomic Mechanisms of Climate Adaptation in Polyploid Bioenergy Switchgrass.” *Nature* 590 (7846): 438–44.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15 (12): 550.

Lü, Peitao, Sheng Yu, Ning Zhu, Yun-Ru Chen, Biyan Zhou, Yu Pan, David Tzeng, et al. 2018. “Genome Encode Analyses Reveal the Basis of Convergent Evolution of Fleshy Fruit Ripening.” *Nature Plants* 4 (10): 784–91.

Lynch, M., and J. S. Conery. 2000. “The Evolutionary Fate and Consequences of Duplicate Genes.” *Science*.

Lyons, Eric, and Michael Freeling. 2008. “How to Usefully Compare Homologous Plant Genes and Chromosomes as DNA Sequences.” *The Plant Journal: For Cell and Molecular Biology* 53 (4): 661–73.

Maere, S., S. De Bodt, J. Raes, T. Casneuf, M. Van Montagu, M. Kuiper, and Y. Van de Peer. 2005. “Modeling Gene and Genome Duplications in Eukaryotes.” *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.0501102102.

Mamidi, Sujan, Adam Healey, Pu Huang, Jane Grimwood, Jerry Jenkins, Kerrie Barry, Avinash Sreedasyam, et al. n.d. “The Setaria Viridis Genome and Diversity Panel Enables Discovery of a Novel Domestication Gene.” https://doi.org/10.1101/744557.

Massey, Frank J. 1951. “The Kolmogorov-Smirnov Test for Goodness of Fit.” *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.1951.10500769.

McCormick, Ryan F., Sandra K. Truong, Avinash Sreedasyam, Jerry Jenkins, Shengqiang Shu, David Sims, Megan Kennedy, et al. 2018. “The Sorghum Bicolor Reference Genome: Improved Assembly, Gene Annotations, a Transcriptome Atlas, and Signatures of Genome Organization.” *The Plant Journal: For Cell and Molecular Biology* 93 (2): 338–54.

Ming, Ray, Shaobin Hou, Yun Feng, Qingyi Yu, Alexandre Dionne-Laporte, Jimmy H. Saw, Pavel Senin, et al. 2008. “The Draft Genome of the Transgenic Tropical Fruit Tree Papaya (Carica Papaya Linnaeus).” *Nature* 452 (7190): 991–96.

Ming, Ray, Robert VanBuren, Yanling Liu, Mei Yang, Yuepeng Han, Lei-Ting Li, Qiong Zhang, et al. 2013. “Genome of the Long-Living Sacred Lotus (Nelumbo Nucifera Gaertn.).” *Genome Biology* 14 (5): R41.

Motamayor, Juan C., Keithanne Mockaitis, Jeremy Schmutz, Niina Haiminen, Donald Livingstone 3rd, Omar Cornejo, Seth D. Findley, et al. 2013. “The Genome Sequence of the Most Widely Cultivated Cacao Type and Its Use to Identify Candidate Genes Regulating Pod Color.” *Genome Biology* 14 (6): r53.

Münkemüller, Tamara, Sébastien Lavergne, Bruno Bzeznik, Stéphane Dray, Thibaut Jombart, Katja Schiffers, and Wilfried Thuiller. 2012. “How to Measure and Test Phylogenetic Signal.” *Methods in Ecology and Evolution*. https://doi.org/10.1111/j.2041-210x.2012.00196.x.

Niederhuth, Chad E., Adam J. Bewick, Lexiang Ji, Magdy S. Alabady, Kyung Do Kim, Qing Li,

Nicholas A. Rohr, et al. 2016. “Widespread Natural Variation of DNA Methylation within Angiosperms.” *Genome Biology* 17 (1): 194.

Niederhuth, Chad E., and Robert J. Schmitz. 2017. “Putting DNA Methylation in Context: From Genomes to Gene Expression in Plants.” *Biochimica et Biophysica Acta, Gene Regulatory Mechanisms* 1860 (1): 149–56.

Noshay, Jaclyn M., Sarah N. Anderson, Peng Zhou, Lexiang Ji, William Ricci, Zefu Lu, Michelle C. Stitzer, et al. 2019. “Monitoring the Interplay between Transposable Element Families and DNA Methylation in Maize.” *PLoS Genetics* 15 (9): e1008291.

Ohno, Susumu. 1970. “Evolution by Gene Duplication.” https://doi.org/10.1007/978-3-642-86659-3.

Ong-Abdullah, Meilina, Jared M. Ordway, Nan Jiang, Siew-Eng Ooi, Sau-Yee Kok, Norashikin Sarpan, Nuraziyan Azimi, et al. 2015. “Loss of Karma Transposon Methylation Underlies the Mantled Somaclonal Variant of Oil Palm.” *Nature* 525 (7570): 533–37.

O’Rourke, Jamie A., Luis P. Iniguez, Fengli Fu, Bruna Bucciarelli, Susan S. Miller, Scott A. Jackson, Philip E. McClean, et al. 2014. “An RNA-Seq Based Gene Expression Atlas of the Common Bean.” *BMC Genomics* 15 (October): 866.

Otto, Sarah P., and Jeannette Whitton. 2000. “Polyploid Incidence and Evolution.” *Annual Review of Genetics*. https://doi.org/10.1146/annurev.genet.34.1.401.

Ou, Shujun, Weija Su, Yi Liao, Kapeel Chougule, Jireh R. A. Agda, Adam J. Hellinga, Carlos Santiago Blanco Lugo, et al. 2019. “Benchmarking Transposable Element Annotation Methods for Creation of a Streamlined, Comprehensive Pipeline.” *Genome Biology* 20 (1): 275.

Pagel, Mark. 1999. “Inferring the Historical Patterns of Biological Evolution.” *Nature*. https://doi.org/10.1038/44766.

Panchy, Nicholas, Melissa D. Lehti-Shiu, and Shin-Han Shiu. 2016. “Evolution of Gene Duplication in Plants.” *Plant Physiology*. https://doi.org/10.1104/pp.16.00523.

Parkin, Isobel A. P., Chushin Koh, Haibao Tang, Stephen J. Robinson, Sateesh Kagale, Wayne E. Clarke, Chris D. Town, et al. 2014. “Transcriptome and Methylome Profiling Reveals Relics of Genome Dominance in the Mesopolyploid Brassica Oleracea.” *Genome Biology* 15 (6): R77.

Paterson, Andrew H., Jonathan F. Wendel, Heidrun Gundlach, Hui Guo, Jerry Jenkins, Dianchuan Jin, Danny Llewellyn, et al. 2012. “Repeated Polyploidization of Gossypium Genomes and the Evolution of Spinnable Cotton Fibres.” *Nature* 492 (7429): 423–27.

Pellicer, Jaume, and Ilia J. Leitch. 2020. “The Plant DNA C□values Database (release 7.1): An Updated Online Repository of Plant Genome Size Data for Comparative Studies.” *New Phytologist*. https://doi.org/10.1111/nph.16261.

Picard, Colette L., and Mary Gehring. 2017. “Proximal Methylation Features Associated with Nonrandom Changes in Gene Body Methylation.” *Genome Biology* 18 (1): 73.

Pophaly, Saurabh D., and Aurélien Tellier. 2015. “Population Level Purifying Selection and Gene Expression Shape Subgenome Evolution in Maize.” *Molecular Biology and Evolution* 32 (12): 3226–35.

Portwood, John L., 2nd, Margaret R. Woodhouse, Ethalinda K. Cannon, Jack M. Gardiner, Lisa C. Harper, Mary L. Schaeffer, Jesse R. Walsh, et al. 2019. “MaizeGDB 2018: The Maize Multi-Genome Genetics and Genomics Database.” *Nucleic Acids Research* 47 (D1): D1146–54.

Qiao, Xin, Qionghou Li, Hao Yin, Kaijie Qi, Leiting Li, Runze Wang, Shaoling Zhang, and

Andrew H. Paterson. 2019. “Gene Duplication and Evolution in Recurring Polyploidization-Diploidization Cycles in Plants.” *Genome Biology* 20 (1): 38.

Raju, Sunil K. Kenchanmane, Sunil K. Kenchanmane Raju, Eleanore Jeanne Ritter, and Chad E. Niederhuth. 2019. “Establishment, Maintenance, and Biological Roles of Non-CG Methylation in Plants.” *Essays in Biochemistry*. https://doi.org/10.1042/ebc20190032.

Revell, Liam J. 2012. “Phytools: An R Package for Phylogenetic Comparative Biology (and Other Things).” *Methods in Ecology and Evolution*. https://doi.org/10.1111/j.2041-210x.2011.00169.x.

Rodin, Sergei N., and Arthur D. Riggs. 2003. “Epigenetic Silencing May Aid Evolution by Gene Duplication.” *Journal of Molecular Evolution* 56 (6): 718–29.

Sato, Shusei, Yasukazu Nakamura, Takakazu Kaneko, Erika Asamizu, Tomohiko Kato, Mitsuteru Nakao, Shigemi Sasamoto, et al. 2008. “Genome Structure of the Legume, Lotus Japonicus.” *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes* 15 (4): 227–39.

Schmutz, Jeremy, Steven B. Cannon, Jessica Schlueter, Jianxin Ma, Therese Mitros, William Nelson, David L. Hyten, et al. 2010. “Genome Sequence of the Palaeopolyploid Soybean.” *Nature* 463 (7278): 178–83.

Schmutz, Jeremy, Phillip E. McClean, Sujan Mamidi, G. Albert Wu, Steven B. Cannon, Jane Grimwood, Jerry Jenkins, et al. 2014. “A Reference Genome for Common Bean and Genome-Wide Analysis of Dual Domestications.” *Nature Genetics* 46 (7): 707–13.

Schnable, James C. 2019. “Genes and Gene Models, an Important Distinction.” *The New Phytologist*, June. https://doi.org/10.1111/nph.16011.

Schultz, Matthew D., Yupeng He, John W. Whitaker, Manoj Hariharan, Eran A. Mukamel, Danny Leung, Nisha Rajagopal, et al. 2015. “Human Body Epigenome Maps Reveal Noncanonical DNA Methylation Variation.” *Nature* 523 (7559): 212–16.

Schultz, Matthew D., Robert J. Schmitz, and Joseph R. Ecker. 2012. “‘Leveling’ the Playing Field for Analyses of Single-Base Resolution DNA Methylomes.” *Trends in Genetics*. https://doi.org/10.1016/j.tig.2012.10.012.

Secco, David, Chuang Wang, Huixia Shou, Matthew D. Schultz, Serge Chiarenza, Laurent Nussaume, Joseph R. Ecker, James Whelan, and Ryan Lister. 2015. “Stress Induced Gene Expression Drives Transient DNA Methylation Changes at Adjacent Repetitive Elements.” *eLife* 4 (July). https://doi.org/10.7554/eLife.09343.

Seymour, Danelle K., Daniel Koenig, Jörg Hagmann, Claude Becker, and Detlef Weigel. 2014. “Evolution of DNA Methylation Patterns in the Brassicaceae Is Driven by Differences in Genome Organization.” *PLoS Genetics*. https://doi.org/10.1371/journal.pgen.1004785.

Sharma, Sanjeev Kumar, Daniel Bolser, Jan de Boer, Mads Sønderkær, Walter Amoros, Martin Federico Carboni, Juan Martín D’Ambrosio, et al. 2013. “Construction of Reference Chromosome-Scale Pseudomolecules for Potato: Integrating the Potato Genome with Genetic and Physical Maps.” *G3* 3 (11): 2031–47.

Silveira, Amanda Bortolini, Charlotte Trontin, Sandra Cortijo, Joan Barau, Luiz Eduardo Vieira Del Bem, Olivier Loudet, Vincent Colot, and Michel Vincentz. 2013. “Extensive Natural Epigenetic Variation at a De Novo Originated Gene.” *PLoS Genetics*. https://doi.org/10.1371/journal.pgen.1003437.

Singh, Rajinder, Meilina Ong-Abdullah, Eng-Ti Leslie Low, Mohamad Arif Abdul Manaf, Rozana Rosli, Rajanaidu Nookiah, Leslie Cheng-Li Ooi, et al. 2013. “Oil Palm Genome Sequence Reveals Divergence of Interfertile Species in Old and New Worlds.” *Nature* 500

(7462): 335–39.
Slotte, Tanja, Khaled M. Hazzouri, J. Arvid Ågren, Daniel Koenig, Florian Maumus, Ya-Long Guo, Kim Steige, et al. 2013. “The Capsella Rubella Genome and the Genomic Consequences of Rapid Mating System Evolution.” *Nature Genetics* 45 (7): 831–35.
Soltis, Pamela S., D. Blaine Marchant, Yves Van de Peer, and Douglas E. Soltis. 2015. “Polyploidy and Genome Evolution in Plants.” *Current Opinion in Genetics & Development* 35 (December): 119–25.
Song, Qingxin, Tianzhen Zhang, David M. Stelly, and Z. Jeffrey Chen. 2017. “Epigenomic and Functional Analyses Reveal Roles of Epialleles in the Loss of Photoperiod Sensitivity during Domestication of Allotetraploid Cottons.” *Genome Biology* 18 (1): 99.
Suyama, Mikita, David Torrents, and Peer Bork. 2006. “PAL2NAL: Robust Conversion of Protein Sequence Alignments into the Corresponding Codon Alignments.” *Nucleic Acids Research* 34 (Web Server issue): W609–12.
Takuno, Shohei, and Brandon S. Gaut. 2012. “Body-Methylated Genes in Arabidopsis Thaliana Are Functionally Important and Evolve Slowly.” *Molecular Biology and Evolution* 29 (1): 219–27.
———. 2013. “Gene Body Methylation Is Conserved between Plant Orthologs and Is of Evolutionary Consequence.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (5): 1797–1802.
Takuno, Shohei, Jin-Hua Ran, and Brandon S. Gaut. 2016. “Evolutionary Patterns of Genic DNA Methylation Vary across Land Plants.” *Nature Plants* 2 (January): 15222.
Tang, Haibao, Vivek Krishnakumar, Shelby Bidwell, Benjamin Rosen, Agnes Chan, Shiguo Zhou, Laurent Gentzbittel, et al. 2014. “An Improved Genome Release (version Mt4.0) for the Model Legume Medicago Truncatula.” *BMC Genomics* 15 (April): 312.
Tran, Robert K., Jorja G. Henikoff, Daniel Zilberman, Renata F. Ditt, Steven E. Jacobsen, and Steven Henikoff. 2005. “DNA Methylation Profiling Identifies CG Methylation Clusters in Arabidopsis Genes.” *Current Biology: CB* 15 (2): 154–59.
Turco, Gina M., Kaisa Kajala, Govindarajan Kunde-Ramamoorthy, Chew-Yee Ngan, Andrew Olson, Shweta Deshphande, Denis Tolkunov, et al. 2017. “DNA Methylation and Gene Expression Regulation Associated with Vascularization in Sorghum Bicolor.” *The New Phytologist* 214 (3): 1213–29.
Tuskan, G. A., S. Difazio, S. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, N. Putnam, et al. 2006. “The Genome of Black Cottonwood, Populus Trichocarpa (Torr. & Gray).” *Science* 313 (5793): 1596–1604.
Valliyodan, Babu, Steven B. Cannon, Philipp E. Bayer, Shengqiang Shu, Anne V. Brown, Longhui Ren, Jerry Jenkins, et al. 2019. “Construction and Comparison of Three Reference-Quality Genome Assemblies for Soybean.” *The Plant Journal: For Cell and Molecular Biology* 100 (5): 1066–82.
VanBuren, Robert, Doug Bryant, Patrick P. Edger, Haibao Tang, Diane Burgess, Dinakar Challabathula, Kristi Spittle, et al. 2015. “Single-Molecule Sequencing of the Desiccation-Tolerant Grass Oropetium Thomaeum.” *Nature* 527 (7579): 508–11.
VanBuren, Robert, Ching Man Wai, Marivi Colle, Jie Wang, Shawn Sullivan, Jill M. Bushakra, Ivan Liachko, et al. 2018. “A near Complete, Chromosome-Scale Assembly of the Black Raspberry (Rubus Occidentalis) Genome.” *GigaScience* 7 (8). https://doi.org/10.1093/gigascience/giy094.
Van de Peer, Yves, Eshchar Mizrachi, and Kathleen Marchal. 2017. “The Evolutionary

Significance of Polyploidy." *Nature Reviews. Genetics* 18 (7): 411–24.

"Various R Programming Tools for Plotting Data [R Package Gplots Version 3.1.1]." 2020, November. https://CRAN.R-project.org/package=gplots.

Verde, Ignazio, Jerry Jenkins, Luca Dondini, Sabrina Micali, Giulia Pagliarani, Elisa Vendramin, Roberta Paris, et al. 2017. "The Peach v2.0 Release: High-Resolution Linkage Mapping and Deep Resequencing Improve Chromosome-Scale Assembly and Contiguity." *BMC Genomics* 18 (1): 225.

Wang, Dapeng, Yubin Zhang, Zhang Zhang, Jiang Zhu, and Jun Yu. 2010. "KaKs_Calculator 2.0: A Toolkit Incorporating Gamma-Series Methods and Sliding Window Strategies." *Genomics, Proteomics & Bioinformatics* 8 (1): 77–80.

Wang, Haifeng, Getu Beyene, Jixian Zhai, Suhua Feng, Noah Fahlgren, Nigel J. Taylor, Rebecca Bart, James C. Carrington, Steven E. Jacobsen, and Israel Ausin. 2015. "CG Gene Body DNA Methylation Changes and Evolution of Duplicated Genes in Cassava." *Proceedings of the National Academy of Sciences of the United States of America* 112 (44): 13729–34.

Wang, Juexin, Md Shakhawat Hossain, Zhen Lyu, Jeremy Schmutz, Gary Stacey, Dong Xu, and Trupti Joshi. 2019. "SoyCSN: Soybean Context□specific Network Analysis and Prediction Based on Tissue□specific Transcriptome Data." *Plant Direct*. https://doi.org/10.1002/pld3.167.

Wang, Jun, Nicholas C. Marowsky, and Chuanzhu Fan. 2014. "Divergence of Gene Body DNA Methylation and Evolution of Plant Duplicate Genes." *PloS One* 9 (10): e110357.

Wang, Lin, Jiahui Xie, Jiantuan Hu, Binyuan Lan, Chenjiang You, Fenglan Li, Zhengjia Wang, and Haifeng Wang. 2018. "Comparative Epigenomics Reveals Evolution of Duplicated Genes in Potato and Tomato." *The Plant Journal: For Cell and Molecular Biology* 93 (3): 460–71.

Wang, W., G. Haberer, H. Gundlach, C. Gläßer, T. Nussbaumer, M. C. Luo, A. Lomsadze, et al. 2014. "The Spirodela Polyrhiza Genome Reveals Insights into Its Neotenous Reduction Fast Growth and Aquatic Lifestyle." *Nature Communications* 5: 3311.

Wang, Xutong, Zhibin Zhang, Tiansi Fu, Lanjuan Hu, Chunming Xu, Lei Gong, Jonathan F. Wendel, and Bao Liu. 2017. "Gene-Body CG Methylation and Divergent Expression of Duplicate Genes in Rice." *Scientific Reports*. https://doi.org/10.1038/s41598-017-02860-4.

Wang, Yupeng, Jingping Li, and Andrew H. Paterson. 2013. "MCScanX-Transposed: Detecting Transposed Gene Duplications Based on Multiple Colinearity Scans." *Bioinformatics* 29 (11): 1458–60.

Wang, Yupeng, Xiyin Wang, Tae-Ho Lee, Shahid Mansoor, and Andrew H. Paterson. 2013. "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*. https://doi.org/10.1111/nph.12137.

Wei, Taiyun, and Simko Viliam. n.d. "R Package 'Corrplot': Visualization of a Correlation Matrix (Version 0.84)." Accessed June 5, 2021. R package "corrplot": Visualization of a Correlation Matrix (Version 0.84). Available from https://github.com/taiyun/corrplot.

Wu, Guohong Albert, Javier Terol, Victoria Ibanez, Antonio López-García, Estela Pérez-Román, Carles Borredá, Concha Domingo, et al. 2018. "Genomics of the Origin and Evolution of Citrus." *Nature* 554 (7692): 311–16.

Xu, Chunming, Brian D. Nadon, Kyung Do Kim, and Scott A. Jackson. 2018. "Genetic and Epigenetic Divergence of Duplicate Genes in Two Legume Species." *Plant, Cell & Environment* 41 (9): 2033–44.

Xue, Huabai, Suke Wang, Jia-Long Yao, Cecilia H. Deng, Long Wang, Yanli Su, Huirong Zhang, et al. 2018. "Chromosome Level High-Density Integrated Genetic Maps Improve the Pyrus Bretschneideri 'DangshanSuli' v1.0 Genome." *BMC Genomics*. https://doi.org/10.1186/s12864-018-5224-6.

Xu, Shuqing, Thomas Brockmöller, Aura Navarro-Quezada, Heiner Kuhl, Klaus Gase, Zhihao Ling, Wenwu Zhou, et al. 2017. "Wild Tobacco Genomes Reveal the Evolution of Nicotine Biosynthesis." *Proceedings of the National Academy of Sciences of the United States of America* 114 (23): 6133–38.

Yanai, Itai, Hila Benjamin, Michael Shmoish, Vered Chalifa-Caspi, Maxim Shklar, Ron Ophir, Arren Bar-Even, et al. 2005. "Genome-Wide Midrange Transcription Profiles Reveal Expression Level Relationships in Human Tissue Specification." *Bioinformatics* 21 (5): 650–59.

Yang, Ruolin, David E. Jarvis, Hao Chen, Mark A. Beilstein, Jane Grimwood, Jerry Jenkins, Shengqiang Shu, et al. 2013. "The Reference Genome of the Halophytic Plant Eutrema Salsugineum." *Frontiers in Plant Science* 4 (March): 46.

Yang, Yu, Kai Tang, Tatsiana U. Datsenka, Wenshan Liu, Suhui Lv, Zhaobo Lang, Xingang Wang, et al. 2019. "Critical Function of DNA Methyltransferase 1 in Tomato Development and Regulation of the DNA Methylome and Transcriptome." *Journal of Integrative Plant Biology* 61 (12): 1224–42.

Zemach, Assaf, Ivy E. McDaniel, Pedro Silva, and Daniel Zilberman. 2010. "Genome-Wide Evolutionary Analysis of Eukaryotic DNA Methylation." *Science* 328 (5980): 916–19.

Zhang, Jianzhi. 2003. "Evolution by Gene Duplication: An Update." *Trends in Ecology & Evolution*. https://doi.org/10.1016/s0169-5347(03)00033-8.

Zhang, Xiaoyu, Junshi Yazaki, Ambika Sundaresan, Shawn Cokus, Simon W-L Chan, Huaming Chen, Ian R. Henderson, et al. 2006. "Genome-Wide High-Resolution Mapping and Functional Analysis of DNA Methylation in Arabidopsis." *Cell* 126 (6): 1189–1201.

Zhao, X. P., Y. Si, R. E. Hanson, C. F. Crane, H. J. Price, D. M. Stelly, J. F. Wendel, and A. H. Paterson. 1998. "Dispersed Repetitive DNA Has Spread to New Genomes since Polyploid Formation in Cotton." *Genome Research* 8 (5): 479–92.

Zheng, Yi, Shan Wu, Yang Bai, Honghe Sun, Chen Jiao, Shaogui Guo, Kun Zhao, et al. 2019. "Cucurbit Genomics Database (CuGenDB): A Central Portal for Comparative and Functional Genomics of Cucurbit Crops." *Nucleic Acids Research* 47 (D1): D1128–36.

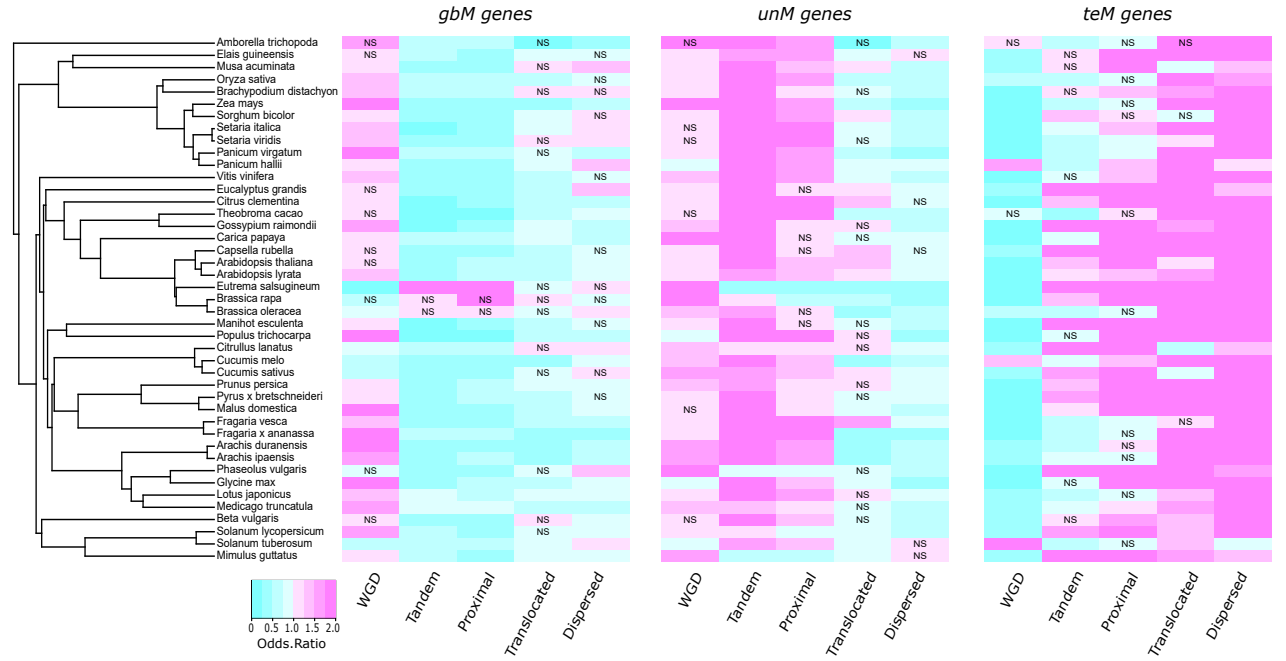


**Figure 1: Patterns of genic methylation across different types of gene duplicates.** Enrichment or depletion of each genic methylation class (gbM, teM, and unM) for each type of gene duplication (WGD, tandem, proximal, translocated, and dispersed). Increasing shades of cyan indicates greater depletion, 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.

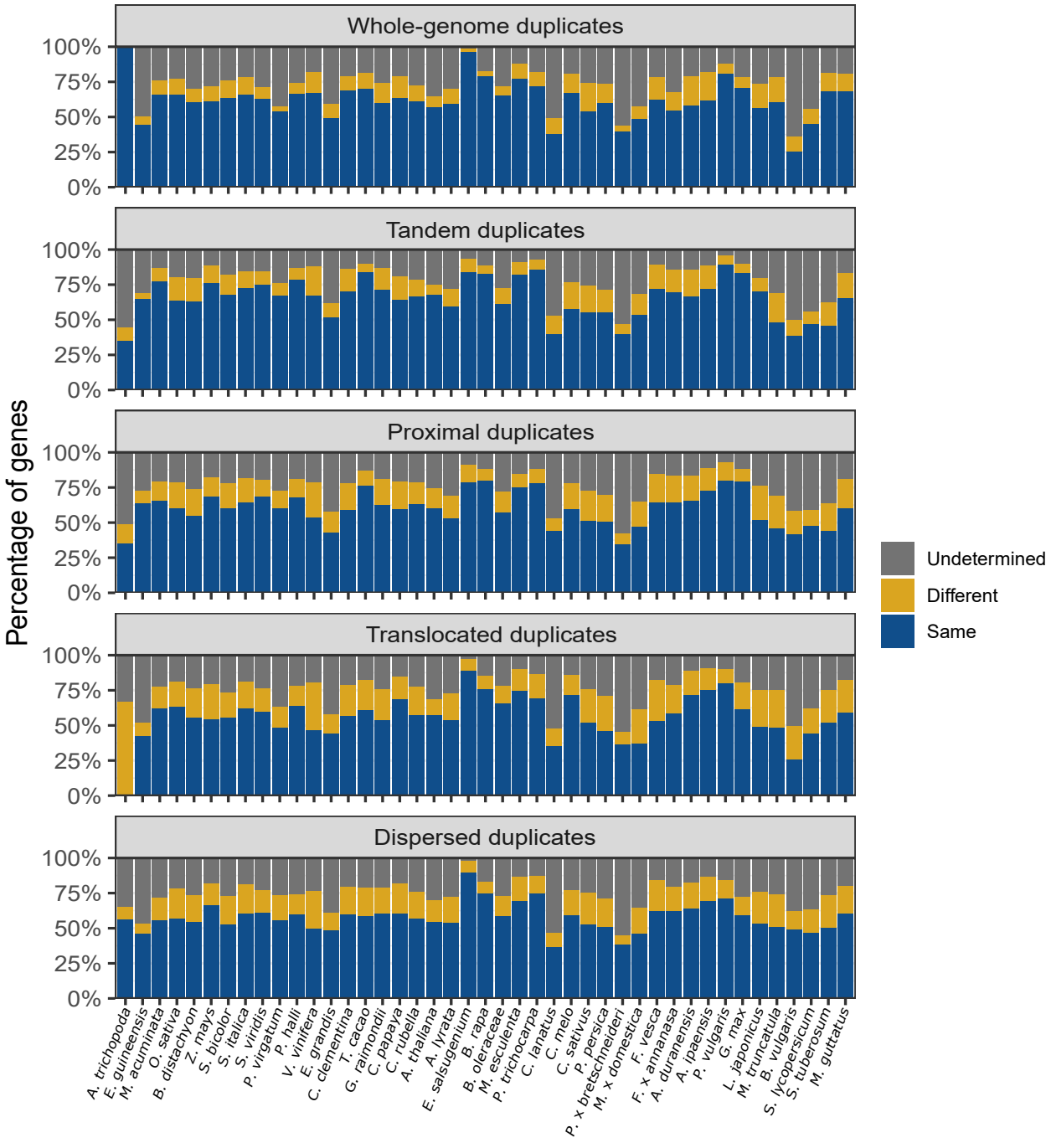


**Figure 2**: **Proportion of paralogs with similar and divergent DNA methylation profiles.**
The proportion of duplicate pairs with similar DNA methylation profiles among different types of duplicate genes (Whole-genome duplicates - WGD, Single-gene duplicates - tandem, proximal, translocated, and dispersed) are shown in Blue. Yellow bars represent 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'.

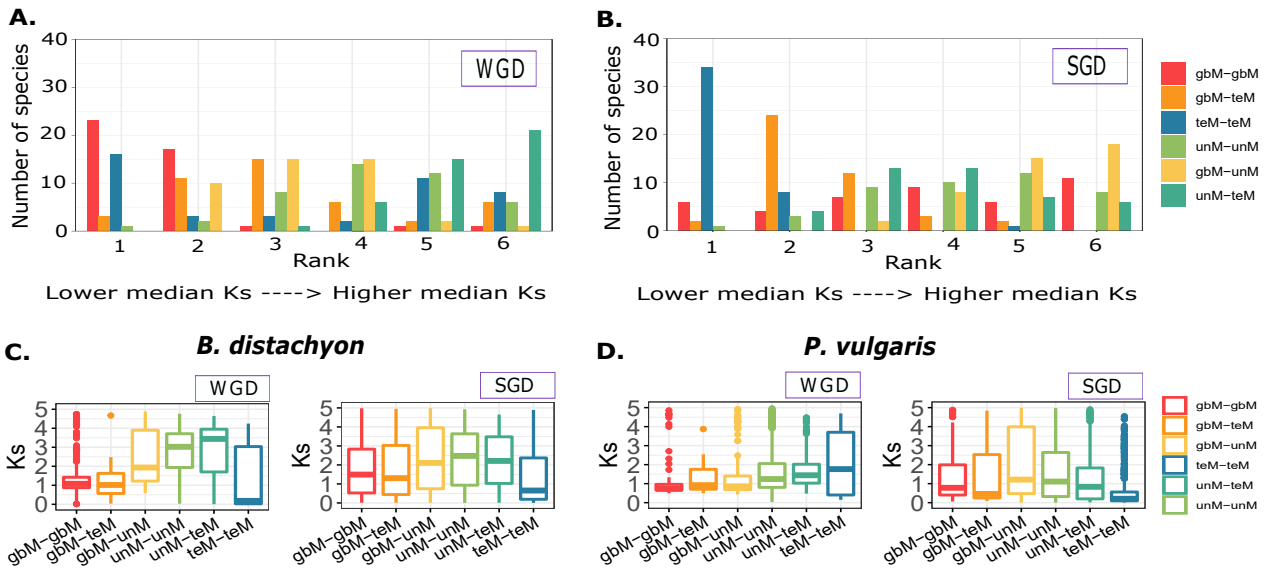


**Figure 3: Relationship between genic methylation and the age of gene duplication.** Bar plots showing the number of species in each of the duplicate-pair genic methylation classifications (gbM-gbM, gbM-teM, teM-teM, unM-unM, gbM-unM, and unM-teM) ranked based on median Ks values (synonymous substitutions) for whole-genome duplicates (A) and single-gene duplicates (B). Box plots (E and F) show the distribution of synonymous substitutions (Ks) for each of the duplicate-pair genic methylation classifications in ***Brachypodium distachyon*** and ***Phaseolus vulgaris*** respectively.

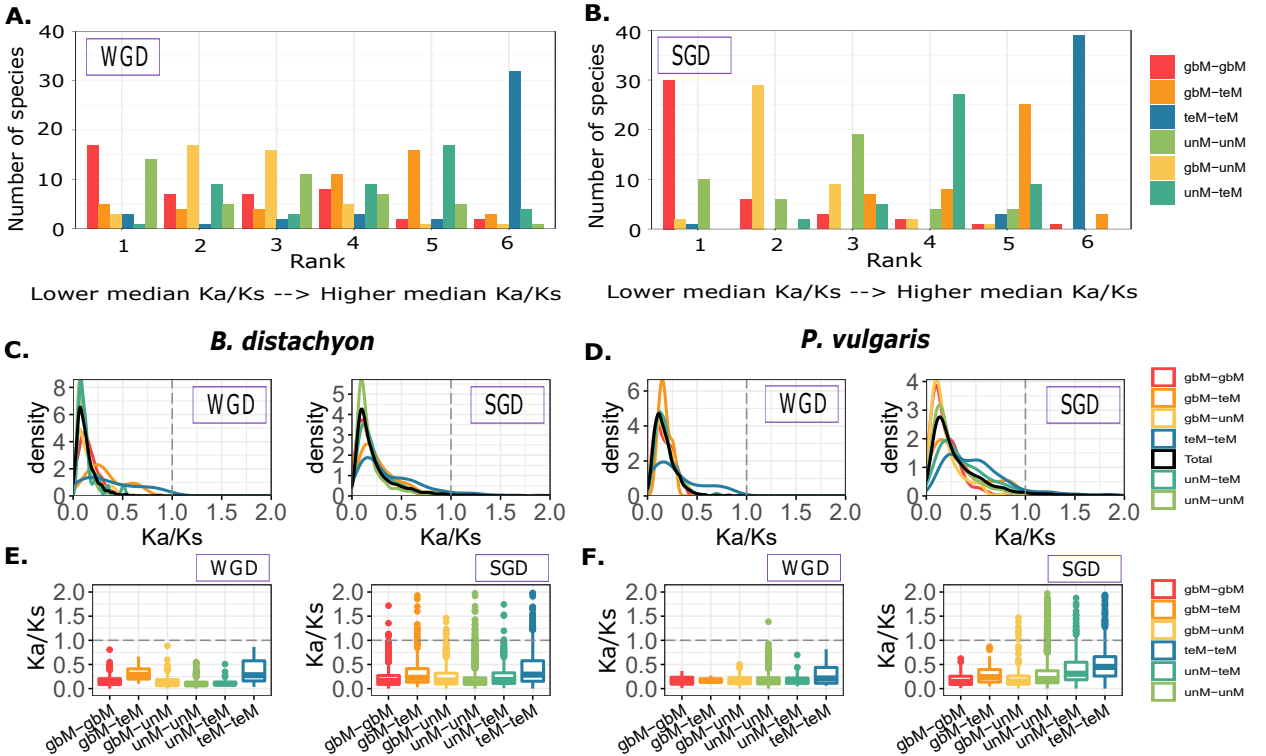


**Figure 4: Relationship between genic methylation and sequence evolution for duplicate pairs.** Bar plots showing the number of species in each of the duplicate-pair genic methylation classifications (gbM-gbM, gbM-teM, teM-teM, unM-unM, gbM-unM, and unM-teM) ranked based on median Ka/Ks values (ratio 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 *Brachypodium distachyon* and *Phaseolus 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.

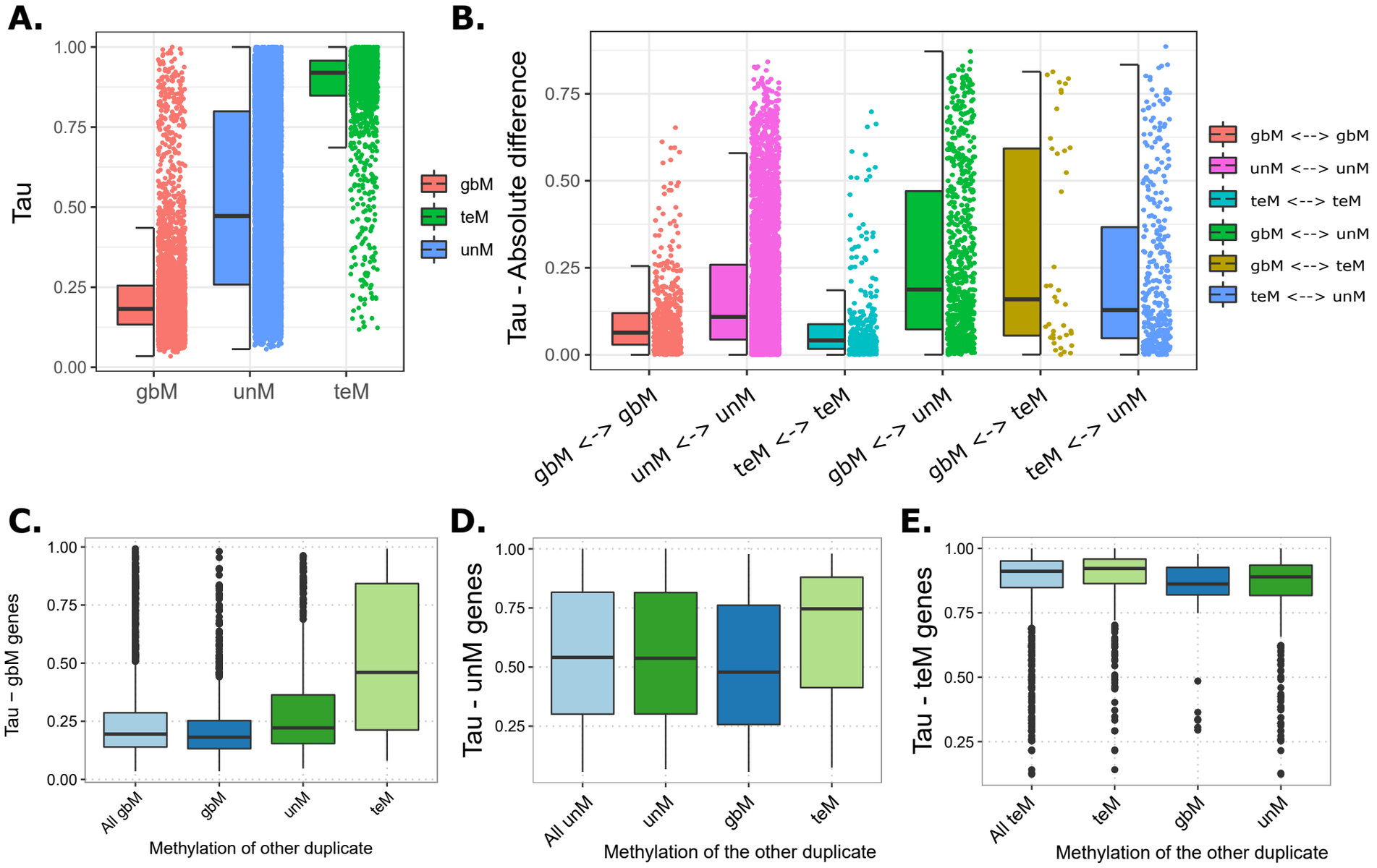


**Figure 5: Gene expression specificity of *A. thaliana* duplicate gene pairs.** Tissue-specificity index, Tau (τ), ranges from 0 (broadly expressed) to 1 (narrowly expressed). (A) Tissue specificity of genes based on genic methylation classification (gbM, unM, and teM). (B) Absolute difference in tissue-specificity 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.

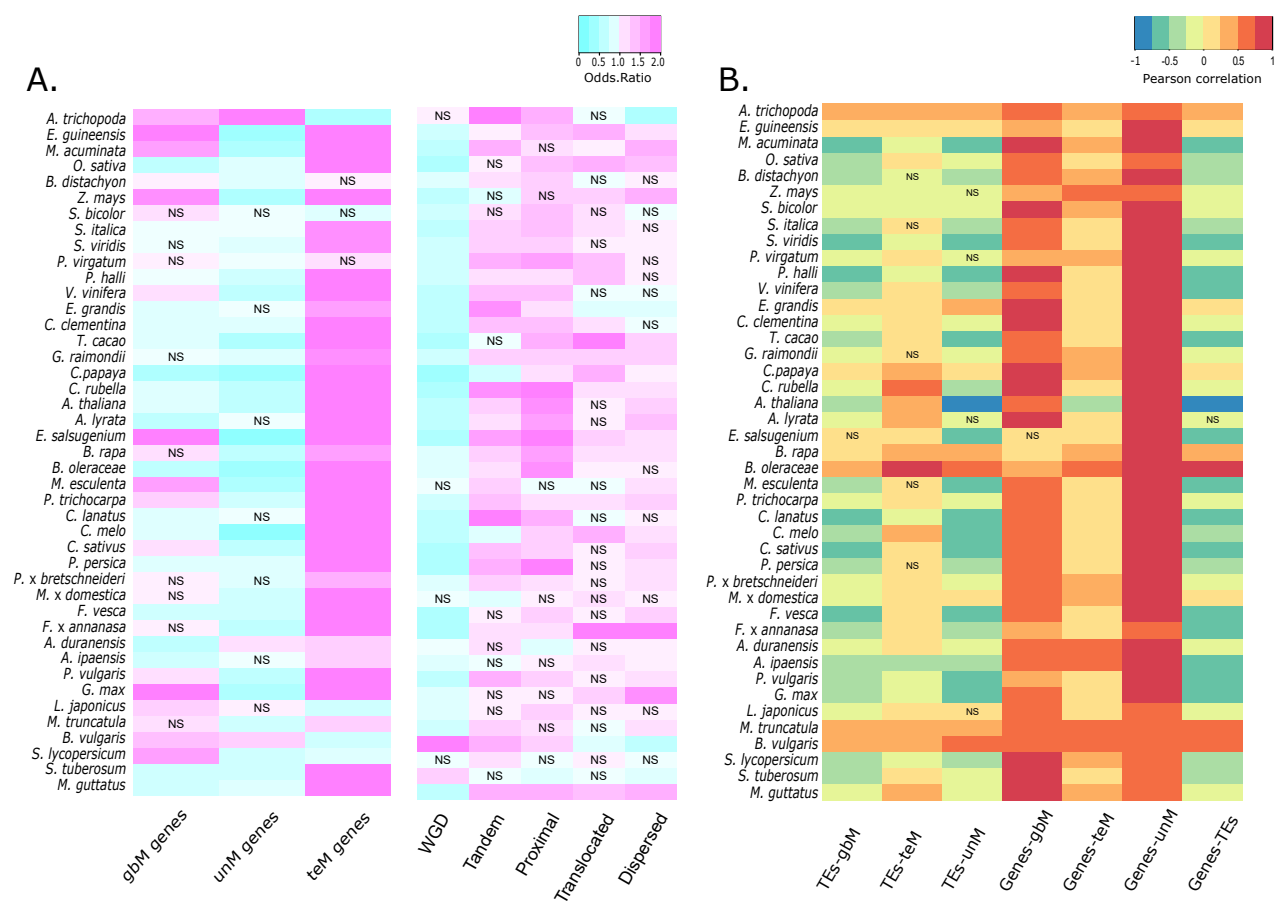


**Figure 6: Local and genome-wide transposon and chromatin environment associations of duplicate genes.**

**A)** Enrichment and depletion of transposable elements (TEs) with gbM, teM, and unM paralogs and different types of duplication in each species. TEs within 1 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.

**B)** Genomic features such as number of genes and number of TEs were calculated in 100kb sliding windows with a 50kb step size. Increasing shades of red indicate positive correlation, while increasing shades of blue represent negative correlations. Unless indicated, all associations are statistically significant at a FDR-corrected p-value < 0.05. 'NS' indicates no statistical significance.

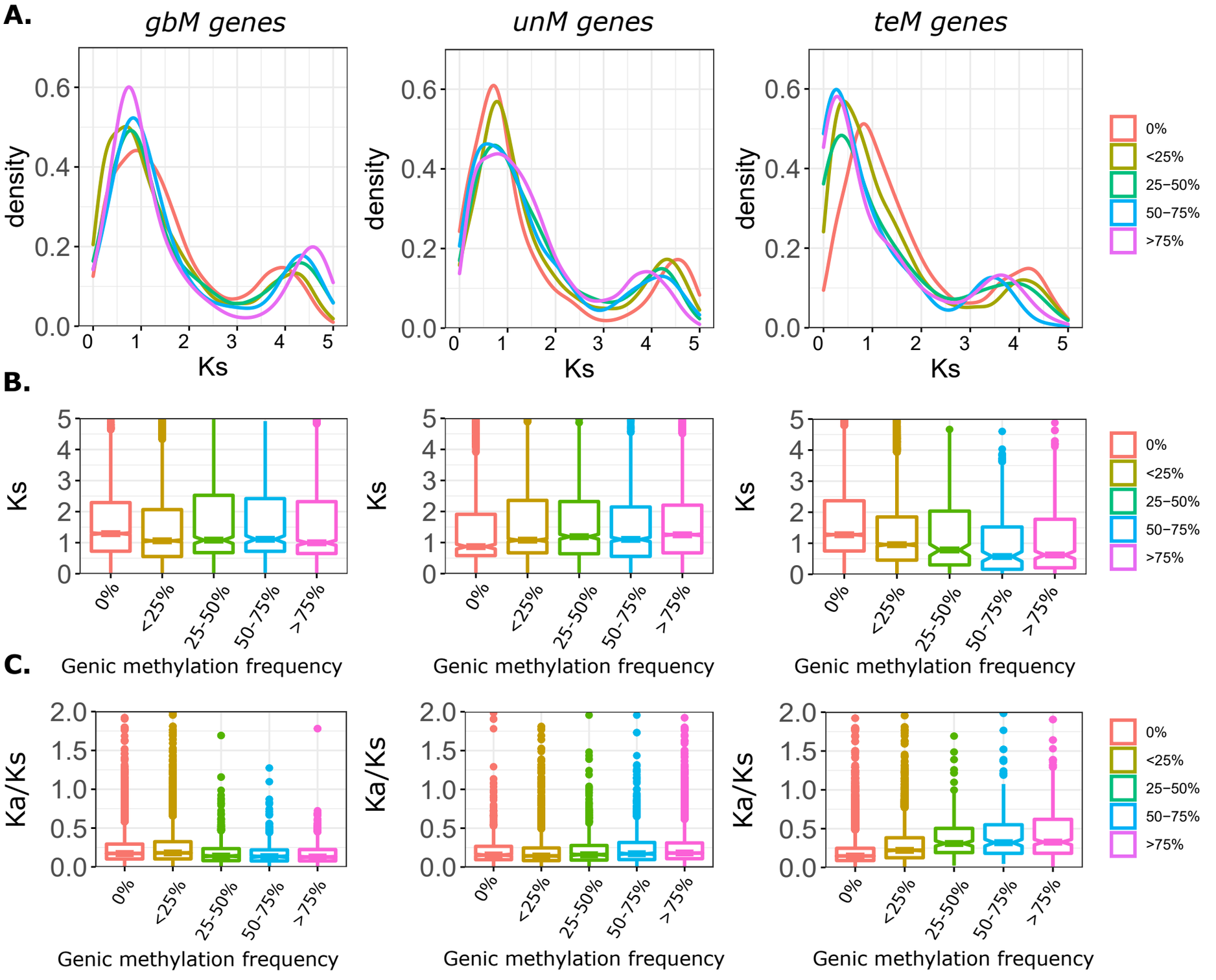


**Figure 7: Genic methylation frequency in a population is associated with age of duplication and sequence evolution**

A) Density plots showing the Ks distribution of genes at different frequencies of gbM, unM, and teM (0%, <25%, 25%-50%, 50%-75%, >75%) in the population. Boxplots of Ks (B) and Ka/Ks distributions (C) for gbM, unM, and teM genes at different frequencies.