




DNA methylation signatures of duplicate gene evolution in angiosperms

Sunil K. Kenchanmane Raju (ಸುನೀಲ ಕೆ. ಕೆಂಚನಮನೆ ರಾಜು) ^{1,*} Marshall Ledford ² and Chad E. Niederhuth ^{1,3,*}

¹ Department of Plant Biology, Michigan State University, East Lansing, MI 48824, USA

² Biology Department, Vassar College, Poughkeepsie, NY 12604, USA

³ AgBioResearch, Michigan State University, East Lansing, MI 48824, USA

*Author for correspondence: niederhu@msu.edu (C.E.N.), kenchanmane@gmail.com (S.K.K.R.)

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (<https://academic.oup.com/plphys/pages/general-instructions>) is: Chad E. Niederhuth (niederhu@msu.edu).

Abstract

Gene duplication is a source of evolutionary novelty. DNA methylation may play a role in the evolution of duplicate genes (paralogs) through its association with gene expression. While this relationship has been examined to varying extents in a few individual species, the generalizability of these results at either a broad phylogenetic scale with species of differing duplication histories or across a population remains unknown. We applied a comparative epigenomic approach to 43 angiosperm species across the phylogeny and a population of 928 *Arabidopsis* (*Arabidopsis thaliana*) accessions, examining the association of DNA methylation with paralog evolution. Genic DNA methylation was differentially associated with duplication type, the age of duplication, sequence evolution, and gene expression. Whole-genome duplicates were typically enriched for CG-only gene body methylated or unmethylated genes, while single-gene duplications were typically enriched for non-CG methylated or unmethylated genes. Non-CG methylation, in particular, was a characteristic of more recent single-gene duplicates. Core angiosperm gene families were differentiated into those which preferentially retain paralogs and “duplication-resistant” families, which convergently reverted to singletons following duplication. Duplication-resistant families that still have paralogous copies were, uncharacteristically for core angiosperm genes, enriched for non-CG methylation. Non-CG methylated paralogs had higher rates of sequence evolution, higher frequency of presence–absence variation, and more limited expression. This suggests that silencing by non-CG methylation may be important to maintaining dosage following duplication and be a precursor to fractionation. Our results indicate that genic methylation marks differing evolutionary trajectories and fates between paralogous genes and have a role in maintaining dosage following duplication.

Introduction

Gene and genome duplication increases organismal gene content, generating a repertoire for functional novelty (Bridges 1935; Ohno 1970; Flagel and Wendel 2009). Whole-genome duplication (WGD) increases the entire gene content (Soltis et al. 2015) and is more pervasive in plants than in other eukaryotes (Otto and Whitton 2000; Van de Peer et al. 2017; Cheng et al. 2018a). Small-scale and single-gene duplications (SGDs) are a continuous

process with ongoing gene birth and death contributing substantially to gene content (Lynch and Conery 2000; Maere et al. 2005; Panchy et al. 2016). The subsequent retention, divergence, and fractionation (loss) of paralogs are biased depending on the duplication type and gene function (Freeling 2009; De Smet et al. 2013).

Factors determining the evolutionary fate of paralogs are an area of intense study, and DNA methylation is thought to be a contributing factor due to its association with gene

expression (Rodin and Riggs 2003; Wang, Wang, et al. 2013). Cytosine methylation at CG dinucleotides is found throughout eukaryotes, while methylation of the non-CG trinucleotide CHG and CHH (H = A, T, or C) contexts is limited to plants (Feng et al. 2010; Zemach et al. 2010). Plant genes have several distinct patterns of DNA methylation within coding regions (henceforth “genetic methylation”) that are associated 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; Zhang et al. 2006) and are frequently conserved between orthologous genes. gbM genes are typically broadly expressed and evolve more slowly (Takuno and Gaut 2012, 2013; Niederhuth et al. 2016; Takuno et al. 2016). Some genes are methylated similar to transposable elements (TEs), having both CG methylation and non-CG methylation within coding regions. This TE-like methylation (teM) is rarely conserved between orthologs and results in transcriptional silencing (Seymour et al. 2014; Niederhuth et al. 2016; El Baidouri et al. 2018). Most genes, however, are unmethylated (unM) and exhibit 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. Tissue-specific silencing of paralogs might lead to “epigenetic complementation,” through subfunctionalization of expression and paralog retention (Adams et al. 2003; Rodin and Riggs 2003). Alternatively, silencing may contribute to pseudogenization and subsequent fractionation (Hua et al. 2013; El Baidouri et al. 2018). Studies in plants (Hua et al. 2013; Wang, Wang, et al. 2013; Wang, Marowsky, and Fan 2014; Kim et al. 2015; Wang et al. 2015, 2017; El Baidouri et al. 2018; Wang et al. 2018; Xu et al. 2018) and animals (Chang and Liao 2012; Keller and Yi 2014) have found that increasing DNA methylation differences between paralogs are associated with divergence in sequence evolution and expression. In soybean (*Glycine max*), gene transposition to heterochromatic regions resulted in silencing by non-CG methylation, increased sequence divergence, and likely pseudogenization (El Baidouri et al. 2018). In the highly duplicated F-box family of *Arabidopsis thaliana*, silencing by DNA methylation and trimethylation of lysine 27 on histone H3 protein (H3K27me3) was associated with increased sequence divergence and was proposed to have a role in maintaining dosage balance (Hua et al. 2013).

Past studies of DNA methylation and gene duplication have been limited to individual species, focused primarily on WGDs, and often ignore the contextual differences of genetic 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 an overarching understanding of the relationship between DNA methylation and paralog evolution. To address these issues, we analyze genetic methylation contexts across 43 angiosperm species and a population of 928 *A. thaliana* ecotypes.

We find overarching trends and relationships between genetic methylation, the type and age of duplication, gene family, and paralog evolution. This work provides a broad phylogenetic and population-scale understanding of the role of DNA methylation in plant duplicate gene evolution and suggests that DNA methylation may have a role in maintaining dosage prior to fractionation.

Results

Genetic methylation across duplication types

We analyzed genetic methylation and gene duplication for 43 angiosperm species (Supplemental Table S1). Genes were classified as gbM, unM, or teM based on DNA methylation in coding regions (Fig. 1A and Supplemental Table S2). Gene duplicates were identified and classified (Supplemental Table S3) as either WGDs or 1 of 4 types of SGD: tandem, proximal, translocated, or dispersed (Fig. 1B). Tandem duplicates occur through unequal crossing-over, resulting in adjacent paralogous copies (Zhang 2003). Proximal duplicates are separated by several intervening genes and arise either through local transposition or interruption of ancient tandem duplicates (Zhao et al. 1998; Freeling et al. 2008). Translocated duplicates (also known as “transposed”) are distally located pairs in which 1 of the genes is syntenic and the other is nonsyntenic (Wang, Li, and Paterson 2013; Qiao et al. 2019) and can arise either by retrotransposition or DNA-based duplication (Cusack and Wolfe 2006). Finally, dispersed duplicates are pairs that fit none of the above criteria and can arise through multiple mechanisms (Ganko et al. 2007; Freeling 2009; Qiao et al. 2019).

We hypothesized that different duplication types would differ in genetic methylation. Each duplication type was tested for enrichment or depletion of gbM, unM, and teM in each species (Fig. 1, C and D, and Supplemental Table S4). Across angiosperms, WGDs were more frequently enriched for gbM (27/43 enriched and 7/43 depleted) and unM (32/43 enriched and 5/43 depleted) and depleted for teM (3/43 enriched and 39/43 depleted). Notable exceptions include 3 Brassicaceae species [cabbage (*Brassica oleracea*), field mustard (*Brassica rapa*), and saltwater cress (*Eutrema salsugineum*)], 3 Cucurbitaceae species [watermelon (*Citrullus lanatus*), muskmelon (*Cucumis melo*), and cucumber (*Cucumis sativus*)], and potato (*Solanum tuberosum*). The 3 Brassicaceae species are known to be depleted of gbM genome wide (Bewick et al. 2016). No known depletion of gbM is documented in the Cucurbitaceae. While *C. melo* WGDs are depleted of gbM and enriched in teM, *S. tuberosum* is the only species showing depletion of both gbM and unM and enrichment of teM in WGDs. “Local” tandem and proximal SGDs are more similar in enrichment/depletion to each other compared with “distal” translocated and dispersed SGDs. Local SGDs are depleted of gbM (tandem and proximal: 40/43 depleted and 1/43 enriched) in all species except for the 3 gbM-deficient Brassicaceae species and are enriched for unM in the majority of species (tandem:

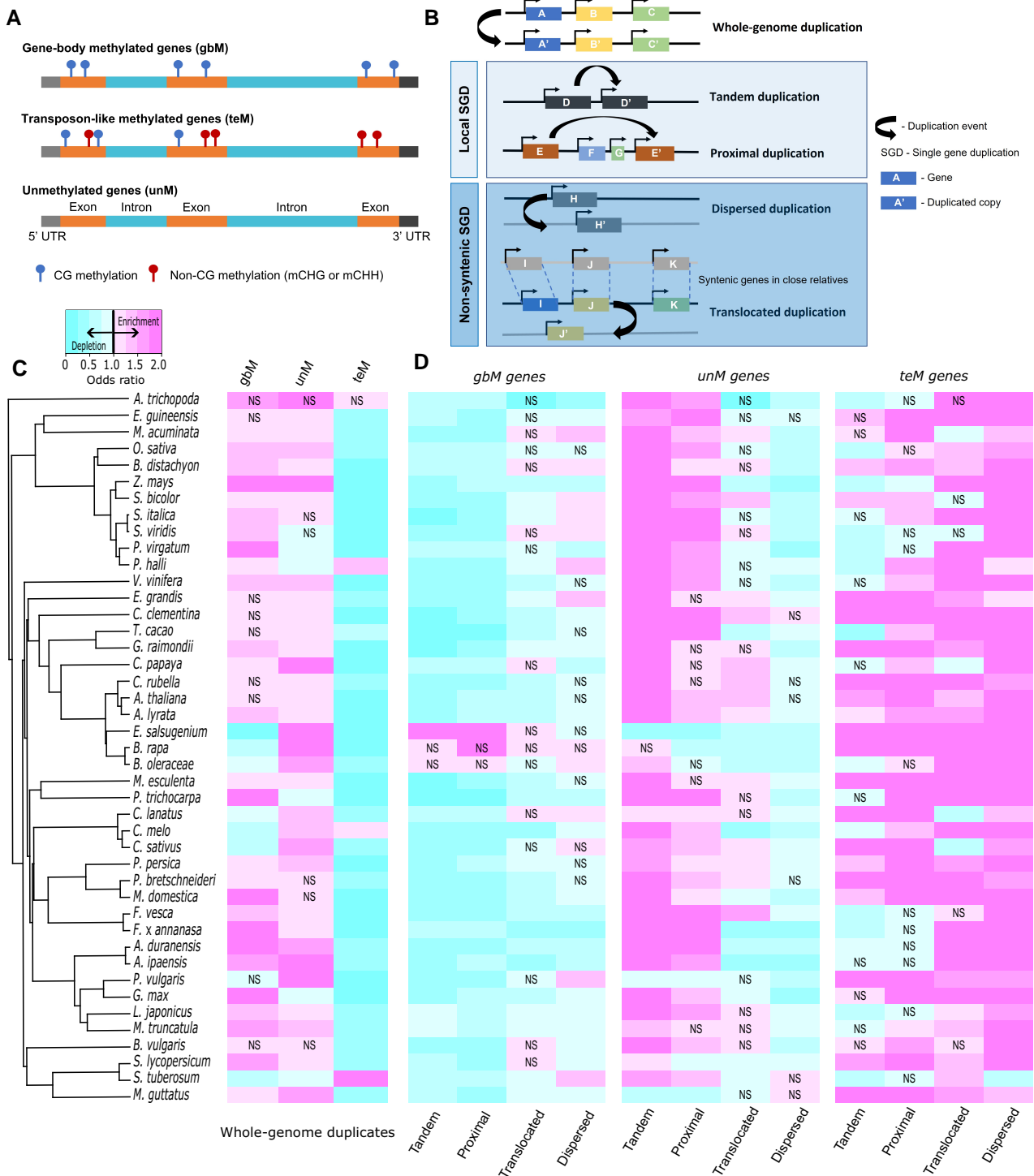


Figure 1. Patterns of genic methylation across different types of gene duplicates. **A**) Schematic representation of genic methylation classification. Coding regions of gbM genes are methylated in the CG context only, teM genes in both CG and non-CG (CHG and CHH) contexts, and unM genes are unmethylated. **B**) Classification of duplicate genes into WGDs and different types of SGDs (tandem, proximal, dispersed, and translocated). **C, D**) Enrichment or depletion of each genic methylation class (gbM, unM, and teM) for **C**) WGDs and **D**) different types of SGDs. A Fisher's exact test odds ratio of <1 represents depletion and >1 indicates enrichment. Unless indicated, all associations are statistically significant at an FDR-corrected $P < 0.05$. "NS" indicates no statistical significance.

39/43 enriched and 3/43 depleted; proximal: 31/43 enriched and 5/43 depleted). Tandem and proximal duplicates differed more in teM (tandem: 19/43 enriched and 14/43

depleted; proximal: 31/43 enriched and 1/43 depleted), with proximal duplicates showing more species enriched for teM than tandem. Like local SGDs, distal SGDs were

more frequently depleted for gbM (translocated: 0/43 enriched and 27/43 depleted; dispersed: 11/43 enriched and 21/43 depleted), although dispersed SGDs were most frequently enriched for gbM. Translocated and dispersed duplicates differ for unM; translocated duplicates have similar numbers of enriched and depleted species (13/43 enriched and 14/43 depleted), while dispersed duplicates are depleted in most species (0/43 enriched and 36/43 depleted). Distal SGDs are more frequently enriched for teM than local SGDs (translocated: 34/43 enriched and 4/43 depleted; dispersed: 42/43 enriched and 1/43 depleted). Increasing teM frequency from tandem to proximal to distal SGDs suggests that teM becomes more common as genes duplicate to increasingly different sequences or chromatin environments.

Effect of gene family on genic methylation and duplication

Gene families differ in their duplicability and retention. Past work has revealed “duplication-resistant” gene families that repeatedly return to single-copy status (Paterson et al. 2006; De Smet et al. 2013; Li et al. 2016), while other gene families retain duplicates over long evolutionary timescales (Conant et al. 2014). To see how gene family composition affects the relationship between genic methylation and duplication, we identified orthogroups for the 43 species with methylome data and additional 15 species included as outgroups (Supplemental Table S1 and Fig. S1). Orthogroups showed a bimodal distribution (Supplemental Fig. S2), with the majority of orthogroups present in either a few species or conserved across most species (Li et al. 2016). Orthogroups represented in ≥ 51 species were classified as “core angiosperm” genes and further divided as “core: single-copy” (duplication-resistant) if represented by a single-copy in $\geq 70\%$ species (Li et al. 2016) and the remainder as “core: multicopy.” The remaining orthogroups were classified based on increasing lineage specificity: “cross family” if present in multiple plant families, “family specific” if restricted to a single family, or “species/lineage specific” if limited to a single species. Genic methylation shows more consistent enrichment and depletion across species for orthogroups than for duplication type (Supplemental Figs. S3 and S4 and Table S5). gbM genes are enriched in core angiosperm orthogroups (multicopy and single-copy) and depleted in noncore orthogroups. The only exceptions were the gbM-depleted species field mustard and saltwater cress. teM genes have the opposite pattern and are depleted in core angiosperm orthogroups and enriched in noncore orthogroups, suggesting a more recent evolutionary origin for gene families enriched for teM. unM genes are variably represented across orthogroups but are more frequently enriched in cross-family (32/43 enriched and 4/43 depleted) and depleted in core: single-copy (2/43 enriched and 41/43 depleted) orthogroups.

We next tested enrichment or depletion of duplicate types in each orthogroup category (Supplemental Fig. S5 and Table S6). SGDs are more frequently enriched in cross-family and family-

specific orthogroups and also in core: multicopy genes, except for proximal duplicates, which is the only duplication type depleted in core: multicopy genes in most species (4/43 enriched and 29/43 depleted). Every duplication type was depleted in core: single-copy orthogroups with the exception of enrichment of dispersed duplicates in soybean and WGDs in strawberry (*Fragaria* \times *ananassa*), soybean, apple (*Malus* \times *domestica*), and switchgrass (*Panicum virgatum*). The greatest enrichment was in the extant polyploids *F.* \times *ananassa* (octoploid) and *P. virgatum* (tetraploid), suggesting insufficient time since WGD to revert to singletons. WGDs were enriched in core: multicopy orthogroups and depleted in noncore orthogroups with the exception of *C. melo* WGDs, which showed enrichment in cross-family orthogroups, and *S. tuberosum* WGDs, which is enriched in cross-family, family-specific, and species/lineage-specific orthogroups. As noted in the previous section, WGDs of these 2 species are depleted for gbM and enriched for teM. The enrichment of WGDs in noncore orthogroups for these species may explain why WGDs differ in their enrichment/depletion of genic methylation. *S. tuberosum* is also the only extant autopolyploid in our data set and is of relatively recent origin (Potato Genome Sequencing Consortium et al. 2011; Wang et al. 2018), which could result in overrepresentation of more lineage-specific genes that are more likely to be teM. Collectively, these results indicate that gene family composition is a driving factor in the relationship between gene duplication and genic methylation.

Methylation divergence between paralogs

Changes in genic methylation might facilitate functional divergence between paralogs and mark different evolutionary trajectories, so we determined the extent of genic methylation differences between paralogs (Fig. 2A and Supplemental Table S7). WGD pairs have the highest similarity in genic methylation (same: $\sim 69\%$ to 97% , median: 84% ; different: $\sim 3\%$ to 31% , median: 16%), followed by tandem (same: $\sim 69\%$ to 93% , median: 82% ; different: $\sim 7\%$ to 31% , median: 18%), proximal (same: $\sim 66\%$ to 90% , median: 77% ; different: $\sim 10\%$ to 34% , median: 23%), and dispersed (same: $\sim 65\%$ to 92% , median: 76% ; different: $\sim 8\%$ to 35% , median: 24%). Translocated duplicates had the broadest range and the greatest proportion of pairs differing in genic methylation (same: $\sim 57\%$ to 90% , median: 74% ; different: $\sim 10\%$ to 42% , median: 25%) (Supplemental Fig. S6 and Table S7). Amborella (*Amborella trichopoda*) is an outlier in this analysis due to the small number of genes classified as WGD or translocated. The direction of genic methylation changes cannot typically be discerned. However, for translocated duplicates, 1 paralog is syntenic and considered parental locus and the translocated gene the daughter locus (Wang, Li, and Paterson 2013). Translocated copies had higher teM proportions in 34/43 species and lower gbM proportions in 24/43 species (Supplemental Table S8). Assuming that parental locus methylation is the original state, we can determine the directionality of methylation changes in the translocated copy. Switching to unM was the most common in 22/43

distributions have been used to date duplication events (Lynch and Conery 2000; Maere et al. 2005). The number and timing of WGD events differ across angiosperms, so we did not expect to find any shared trends for WGD Ks. Contrary to this expectation, WGD gbM-containing pairs tended to have lower Ks values and unM-containing pairs higher Ks values (Fig. 3A and Supplemental Table S11). No clear trend was observed for teM-containing WGD pairs. This may be due to the depletion and lower numbers of teM in WGDs. In contrast to WGD, SGD is a continuous process with constant gene birth and death (Lynch 2002). SGD teM-containing pairs typically had lower Ks values than those with only gbM or unM paralogs (Figs. 3B and S7). This is most evident for teM–teM pairs, but gbM–teM and unM–teM also have lower Ks values. This suggests that teM paralogs tend to be evolutionarily younger. We confirmed this using a method independent of Ks for translocated genes. As the syntenic gene is assumed to be parental in translocated genes, the daughter gene can be parsed into different periods (epochs) at each node of the species tree (Supplemental Table S12) by sequential exclusion to the closest outgroup (Wang, Li, and Paterson 2013). More recent translocated duplicates were enriched in teM paralogs, while more ancient translocated duplicates were enriched for gbM and unM paralogs (Supplemental Fig. S8 and Table S13), supporting our observations from the Ks analysis. These results also fit with the observation that evolutionarily younger lineage-specific orthogroups are enriched for teM genes. We also compared Ks distributions for SC-intermediates and core:multicopy genes but observed no difference suggesting similar evolutionary ages (Supplemental Fig. S9).

The ratio of nonsynonymous (Ka) to synonymous (Ks) substitutions (Ka/Ks) is indicative of sequence evolution (Miyata and Yasunaga 1980; Yang and Bielawski 2000). A Ka/Ks < 1 is indicative of purifying selection, Ka/Ks = 0 indicates neutral selection, and Ka/Ks > 1 is indicative of diversifying selection. We calculated Ka/Ks ratios for each duplicate pair and examined their distributions (Supplemental Fig. S10 and Table S14). The majority of pairs have a Ka/Ks < 1, regardless of duplication type or genic methylation, indicating purifying selection. However, there are differences in the distribution based on genic methylation. For both WGD and SGD genes (Fig. 3, C and D), gbM-containing pairs have lower Ka/Ks, teM-containing pairs have higher Ka/Ks, while unM-containing pairs are intermediate in distribution. This suggests that teM paralogs are under relaxed selective constraints compared with gbM and unM paralogs. A number of pairs had a Ka/Ks > 1 indicating diversifying selection. These were enriched for SGD and depleted for WGD in almost every species except *S. tuberosum* (Supplemental Table S15) and enriched in teM-containing pairs, in particular teM–teM pairs (Supplemental Table S16). We hypothesized that SC-intermediates would be under relaxed selective constraints as these are enriched for teM. Indeed, SC-intermediates have higher Ka/Ks values compared with core:multicopy pairs (Supplemental Fig. S11). Increased

nonsynonymous substitutions in SC-intermediates could lead to their pseudogenization and facilitate fractionation to singleton status.

Ongoing gene duplication and differential fractionation within a species can create presence–absence variation (PAV). We used published lists of PAVs in *B. oleracea*, maize (*Zea mays*), tomato (*Solanum lycopersicum*), and *S. tuberosum* (Hirsch et al. 2014; Golicz et al. 2016; Hardigan et al. 2016; Gao et al. 2019) to examine the relationship between PAVs, genic methylation, and gene duplication (Supplemental Fig. S12 and Table S17). PAVs are enriched for teM genes in all 4 species and depleted for gbM in 3 species, except gbM-deficient *B. oleracea*. PAVs are depleted for unM genes in 3 species and enriched for unM in *S. lycopersicum*. Results are identical whether tested for all genes or duplicated genes only. Association between PAVs and teM could result from targeting of lineage-specific SGDs or incomplete fractionation of teM duplicates in the population. An example of the latter would be fractionation of core: single-copy orthogroups following WGD. In all 4 species, SC-intermediates had higher frequencies of PAV compared with core:multicopy orthogroups and a higher frequency compared with all genes in maize (Supplemental Fig. S13), supporting the hypothesis that teM silencing is an intermediate to fractionation for single-copy core angiosperm genes.

Genic methylation and paralog expression

Divergent expression between paralogs is proposed to be the first step in functional diversification, enabling paralogs to subfunctionalize in expression and increasing the odds of retention (Ohno 1970; Ferris and Whitt 1979; Li et al. 2005). We used gene expression atlases in *A. thaliana*, *G. max*, common bean (*Phaseolus vulgaris*), and sorghum (*Sorghum bicolor*) (O'Rourke et al. 2014; Klepikova et al. 2016; McCormick et al. 2018; Wang et al. 2019) to explore the relationship between genic methylation and paralog expression. For each gene, we calculated the expression specificity (τ), a measure of the number of conditions in which a gene is expressed (Yanai et al. 2005). The value of τ ranges from “0” (broad expression) to “1” (narrow expression). Typically, gbM genes have the broadest expression, teM the narrowest, while unM genes have a wide range of expression specificity (Figs. 4A and S14). Core angiosperm genes are more broadly expressed, expression specificity becoming narrower with increasing lineage specificity (Supplemental Fig. S15). This trend persists when broken down by genic methylation. gbM, unM, and teM genes become more narrowly expressed with increasing lineage specificity. By duplication type, WGDs have broader expression and SGDs narrower expression (Supplemental Fig. S16). Distal SGDs have broader expressions than local SGDs, even though distal SGDs can place duplicates in new chromatin contexts and are more frequently enriched for teM. Similar to what was observed for orthogroups, differences between duplication types persist even when comparing genes of like genic methylation (Supplemental Fig. S16). WGD teM genes had overall broader

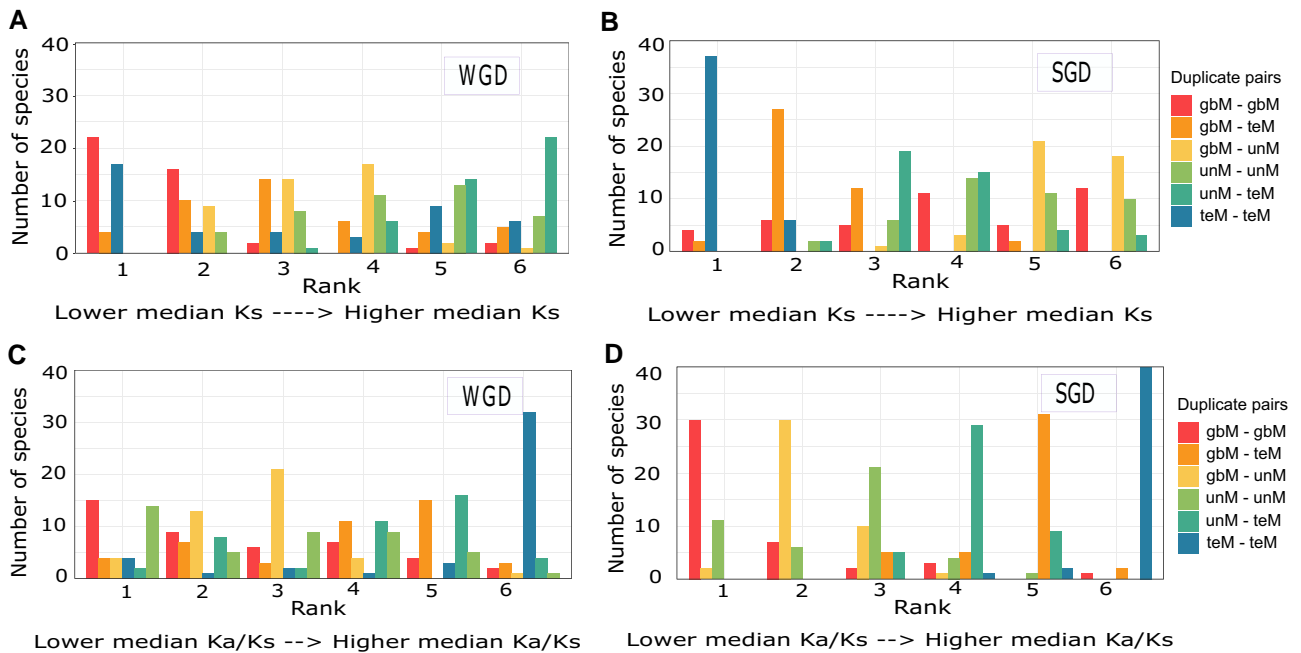


Figure 3. Relationship between genic methylation, age of duplication, and sequence evolution of duplicate paralogs. The number of species in each of the duplicate pair genic methylation classifications (gbM–gbM, gbM–teM, teM–teM, unM–unM, gbM–unM, and unM–teM) ranked based on median Ks values (synonymous substitutions) for WGDs **A**) and SGDs **B**). 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 WGDs **C**) and SGDs **D**).

expression than distal teM SGD teM genes, which had broader expression than local SGD teM genes. This same trend was observed for gbM and unM. Both orthogroup and duplication type therefore exert an effect beyond the genic methylation, and global differences between both orthogroups and duplication types are not fully explained by differential enrichment of genic methylation.

We next examined expression divergence between duplicate pairs, first calculating the expression correlation of each pair (Supplemental Fig. S17). Pairs with the same genic methylation had higher correlation than pairs differing in genic methylation, gbM–gbM pairs having the highest overall correlation. Fitting this, duplicate pairs differing in genic methylation have a greater overall absolute difference in expression specificity than pairs with the same genic methylation (Figs. 4B and S18 and Supplemental Table S18). Surprisingly, the expression specificity differed for genes of the same genic methylation based on the methylation of its duplicate pair (Figs. 4, C to E, and S19). gbM genes that are part of gbM–gbM pairs tend to have broader expression specificity compared with the gbM genes in gbM–unM and gbM–teM pairs. For unM genes, those in unM–gbM pairs had broader expression specificity than those in unM–unM pairs, and those in unM–teM pairs had narrower expression specificity than either. Finally, teM genes in teM–gbM pairs had broader expression specificity than those in teM–unM pairs, which in turn had broader expression specificity than those in teM–teM pairs. This suggests a potential relationship between the parental locus expression and the

expression of the duplicate copy. Perhaps, certain genes may be predisposed to genic methylation changes by their expression.

Transposons and chromatin environment associations

Non-CG methylation is often associated with Transposons (TEs), and TEs can alter the gene chromatin and expression (Hirsch and Springer 2017; Raju et al. 2019). We identified TEs in or within 1 kb (Fig. 5A and Supplemental Table S19) for each paralog. teM paralogs are enriched (36/43 enriched and 4/43 depleted) and unM paralogs are depleted (3/43 enriched and 34/43 depleted) for TEs in the majority of species. gbM was enriched (15/43) and depleted (15/43) for TEs in an equal number of species. Examining duplication type (Fig. 5B and Supplemental Table S20), WGDs are depleted (2/43 enriched and 37/43 depleted) and all 4 SGDs enriched (tandem: 30/43 enriched and 3/43 depleted; proximal: 33/43 enriched and 2/43 depleted; translocated: 21/43 enriched and 2/43 depleted; dispersed: 27/43 enriched and 4/43 depleted) for TEs in the majority of species. Enrichment of TEs in SGDs may partly explain enrichment of teM in SGDs. We compared TE presence/absence for duplicate pairs differing in genic methylation (Supplemental Table S21), hypothesizing that the teM paralog would associate with TEs more frequently than its unM or gbM pair. This was true for *A. thaliana* and lirate rockcress (*Arabidopsis lyrata*), but for most species, both paralogs are associated with a TE in gbM–teM

and unM–teM pairs, suggesting a more complex relationship than simple TE presence/absence in switching of genic methylation states. Differences in TE location or the TE family may be relevant, as was shown with TEs and heterochromatin spreading in *Z. mays* (Eichten et al. 2012).

Location and chromatin environment can also affect genic methylation of duplicate genes. In *G. max*, translocation of paralogs to TE-rich pericentromeric regions often resulted in teM acquisition (El Baidouri et al. 2018). We used gene number, TE number, and fraction of TE base pairs in sliding windows across the genome as a proxy for regions of euchromatin and heterochromatin and correlated these with the number of gbM/unM/teM paralogs (Fig. 5C and Supplemental Table S22). gbM, unM, and teM duplicates are positively correlated with gene number, except *A. thaliana*, where teM has a very weak negative correlation [Pearson's $r = -0.05$, false discovery rate (FDR)-corrected $P = 0.004$]. This may be due to *A. thaliana* genomic organization, which has the smallest genome and the strongest negative correlation between total gene number and TEs (Pearson's $r = -0.71$, FDR-corrected $P < 0.001$) in our data. In most species, the distribution of gbM and unM genes is negatively correlated (gbM: 8/43 positive and 33/43 negative; unM: 10/43 positive and 31/43 negative) and teM genes positively correlated with TEs (24/43 positive and 8/43 negative). This supports the hypothesis that duplication of genes to heterochromatic regions can lead to teM acquisition; however, this does not explain cases such as SC-intermediates.

Epiallele frequency and paralog evolution within a population

DNA methylation varies across a population (Becker and Weigel 2012). The relationship between this variation and paralog evolution is unknown. To address this, genes were classified based on genic methylation and binned according to the frequency of gbM/unM/teM across 928 *A. thaliana* accessions (Kawakatsu et al. 2016). We examined the proportion of duplication type (Fig. 6A) and orthogroup (Supplemental Fig. S20) for each bin, predicting that there would be a corresponding change according to the frequency of genic methylation. This was true only in some instances. gbM showed a slight decrease in SGDs at higher frequencies, this being the most evident for tandem SGDs, while tandem SGDs continually increased with unM frequency. WGD decreased and proximal SGDs increased with increasing teM frequency, while tandem SGDs peaked at ~25% to 50% before declining. As observed across species, a stronger association was observed for orthogroups. As the gbM frequency increases, the frequency of core:multicopy orthogroups increases, and the frequency of cross-family, family-specific, and species/lineage-specific orthogroups decreases. We observed an opposite trend with increasing teM frequency. As the unM frequency increases in the population, we observe an increase in cross-family orthogroups and a decrease in the frequency of core single-copy genes.

We examined Ks (Fig. 6B) and Ka/Ks (Fig. 6C) at different epiallele frequencies. Both gbM and unM show little variation in Ks at different frequencies, while Ks steadily decreases with increasing teM frequency. This suggests a higher frequency of teM in evolutionarily more recent paralogs. Ka/Ks decreased with increasing gbM frequency and increased with increasing teM frequency, fitting the expectation of gbM genes being under greater purifying selection and teM genes being under relaxed selective constraints. Ka/Ks increased slightly with higher unM frequency. We hypothesized that this may be related to gene expression. Supporting this hypothesis, we observed that τ increased (more tissue specific) at higher population frequencies of both unM and teM and decreased with increasing frequency of gbM (Fig. 6D).

Discussion

DNA methylation has been proposed to have a role in paralog evolution (Rodin and Riggs 2003; Wang, Wang, et al. 2013; Keller and Yi 2014; Wang, Marowsky, and Fan 2014). However, this has not been examined at either a broad phylogenetic level or within a population, leaving the generalizability of results from individual species unresolved. To address this, we examined DNA methylation and paralog evolution across 43 angiosperms and a population of 928 *A. thaliana* accessions. Across the phylogeny, WGDs are broadly enriched for gbM and unM genes and depleted for teM genes. There is further differentiation between “local” SGDs (tandem and proximal) and “distal” SGDs (translocated and dispersed). Both are more frequently depleted in gbM, local duplicates are more frequently enriched for unM, while there is an increasing frequency of teM from tandem to proximal to translocated to dispersed SGDs. There are notable exceptions to these trends. For 3 Brassicaceae species (*B. oleracea*, *B. rapa*, and *E. salsugineum*), divergence from these patterns is explained by a known depletion of gbM (Bewick et al. 2016). The Cucurbitaceae and *S. tuberosum* were also depleted of WGD gbM, despite no known depletion of gbM in these species, and need further investigation. We observe an even more consistent association of genic methylation with different types of orthogroups, and this may drive patterns observed between genic methylation and gene duplication. This appears to be the case for *S. tuberosum*, a relatively recent autopolyploid (Potato Genome Sequencing Consortium et al. 2011). Unlike other species, *S. tuberosum* WGDs are enriched for increasingly lineage-specific orthogroups which explains the depletion of gbM and unM and enrichment of teM in WGDs.

gbM is characteristic of evolutionarily conserved genes (Takuno and Gaut 2012, 2013; Bewick et al. 2016; Takuno et al. 2016). Fitting this, gbM is enriched in core angiosperm genes, and gbM–gbM duplicate pairs are more conserved in sequence and expression. In contrast, teM genes are evolutionarily younger, increasing in enrichment with greater lineage specificity, and predominantly found in recent SGDs. teM paralogs have narrower expression, higher Ka/Ks ratios,

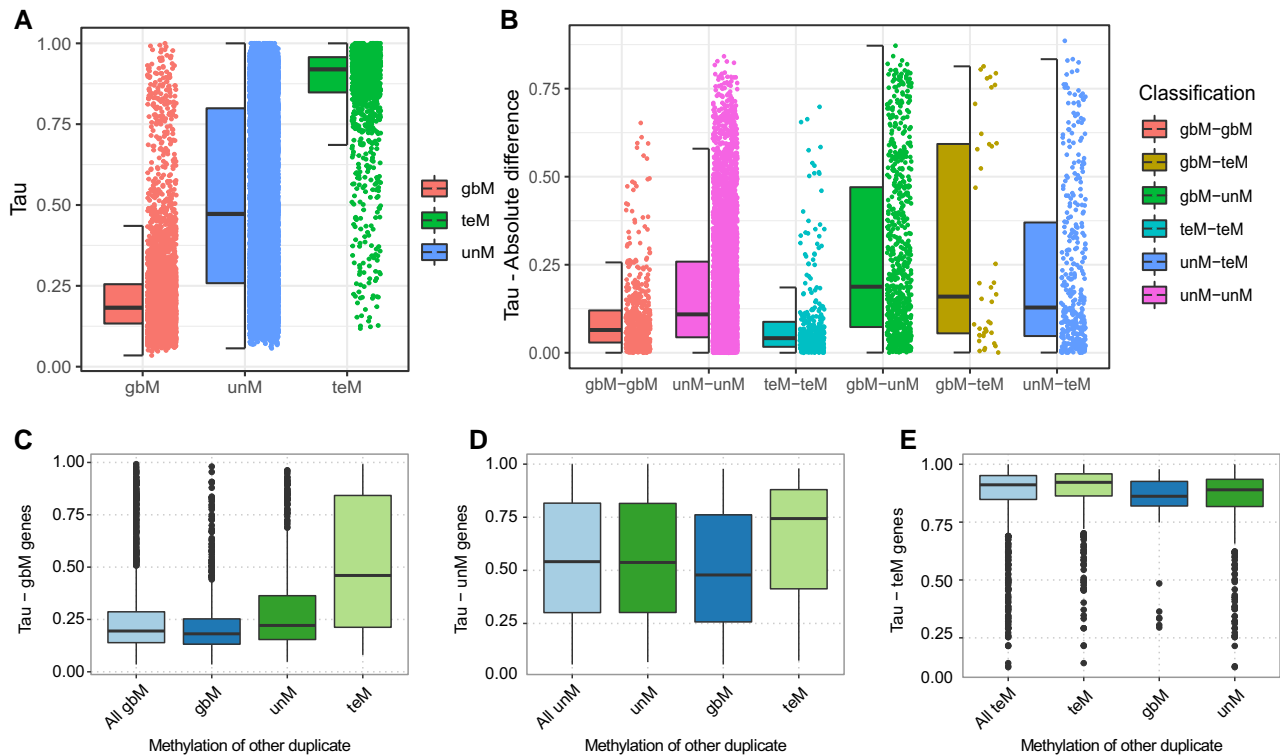


Figure 4. Gene expression specificity of *A. thaliana* duplicate genes. Tissue specificity index (τ), ranges from 0 (broadly expressed) to 1 (narrowly expressed). **A)** Tissue specificity of genes based on genic methylation classification (gbM, unM, and teM). Each dot represents τ value of a gene. Center line in the half boxplot represents the median τ while the box limits represent 25% and 75% percentile of the interquartile range; whiskers represent 1.5 times above or below the interquartile range. **B)** Absolute difference in tissue specificity index (τ) between pairs of duplicate genes with similar or divergent methylation. Tissue specificity index (τ) of **C)** gbM, **D)** unM, and **E)** teM genes when the other duplicate pair has the same or a different genic methylation status. For example, for **C)** gbM genes τ was plotted for all gbM genes and the gbM paralog in gbM–gbM, gbM–teM, and gbM–unM pairs. Similarly, τ of only the unM paralog is plotted for **D)** unM genes and τ of the teM paralog for **E)** teM genes. Dots represent outliers 1.5 times above or below the interquartile range.

and enrichment in PAV, suggesting that most are on the path to pseudogenization. This process would lead to their depletion in more ancient WGDs, while continual duplication in SGDs would result in their enrichment. unM genes are seemingly intermediate between gbM and teM in most aspects. unM might be considered the “default” state and spans from more gbM-like to more teM-like genes. In gbM-depleted species, the gbM ortholog is unM (Bewick et al. 2016). unM is the largest group and broadly represented across both core angiosperm orthogroups and more lineage-specific orthogroups. Many transcription factors and kinases are retained following WGD and have tissue-specific expression characteristic of unM (Pophaly and Tellier 2015). At the same time, many tandem and proximal duplications are associated with environmental adaptation (Freeling 2009). This would favor retention of unM in both WGDs and local SGDs. Unexpectedly, unM-containing WGD pairs typically have a higher Ks than gbM-containing pairs. We speculate that this could result from unM-containing pairs being derived from more ancient WGD events or differences in the mutation rates of gbM and unM genes.

Within a population, paralog evolution is associated with genic methylation frequency. This is especially true for teM

genes, where increasing teM frequency is associated with evolutionary younger genes, narrower expression, and greater sequence divergence. Differences are also observed in gbM and unM genes and appear to be driven by expression as more narrowly expressed unM genes have a higher Ka/Ks ratio. To achieve high frequency in a population, the simplest explanation is that a genic methylation state was established early following duplication and spread with expansion of the population. Low-frequency states would reflect either relict populations (Kawakatsu et al. 2016) or cases of recent acquisition. A deeper analysis taking into account the structure and relationship of accessions in the population will provide further insight into the establishment of genic methylation states and paralog evolution.

Gene family is also an important factor, as indicated by our orthogroup analyses, and should be accounted for when trying to understand the role of DNA methylation in paralog evolution. Gene families differ in susceptibility to fractionation, some being preferentially retained, while others convergently revert to singletons. Both are thought to be dosage sensitive. The former retains duplicate copies to maintain relative dosage to other genes in the genome as explained by the gene balance hypothesis (Birchler and Veitia 2010)

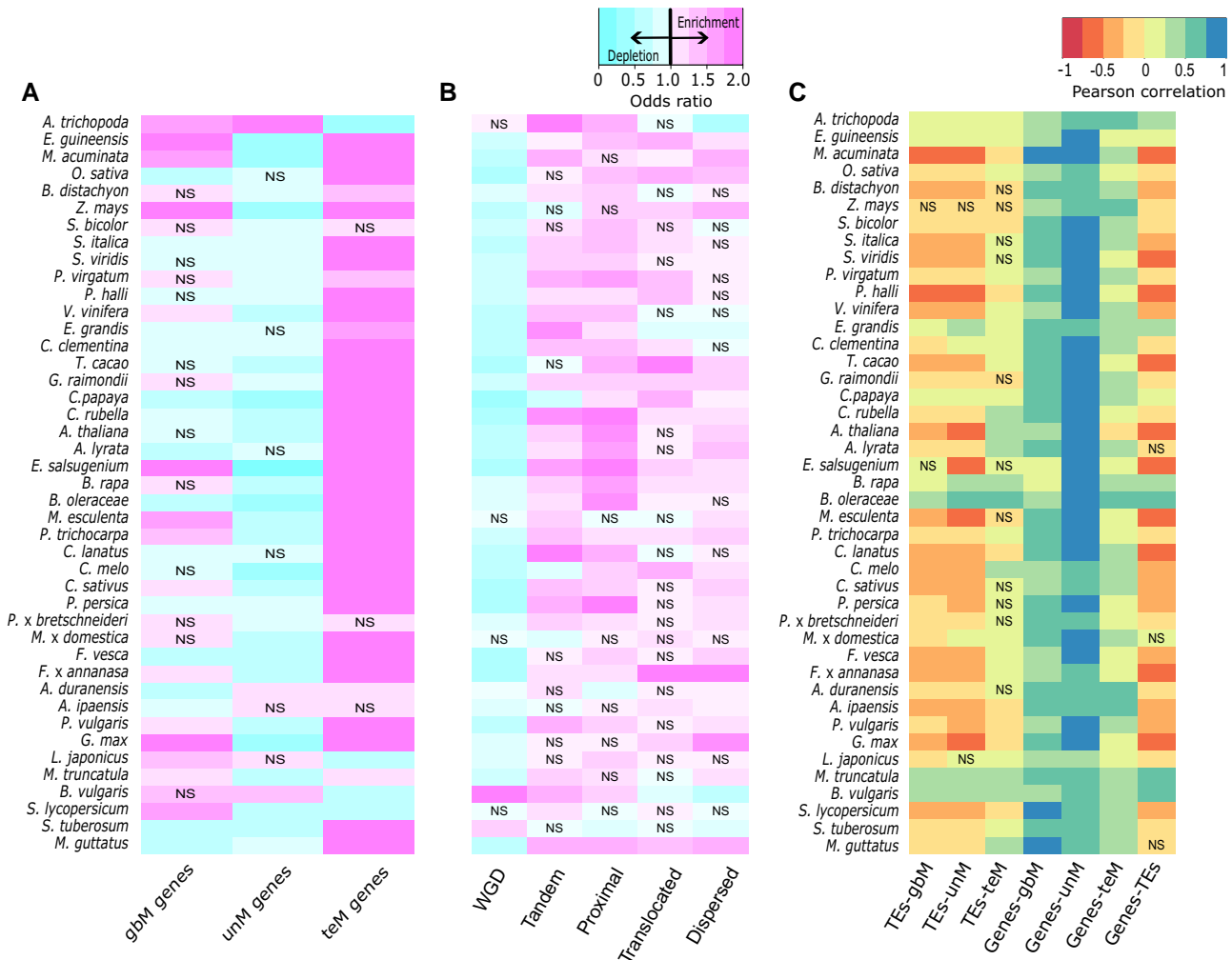


Figure 5. Association of TEs and chromatin environment with duplicate genes. Enrichment and depletion of TEs with duplicate genes based on **A)** genic methylation classification and **B)** type of duplication in each species. TEs within 1-kb upstream, downstream, or within the gene body were considered associated with that gene. A Fisher's exact test odds ratio of <1 represents depletion and >1 indicates enrichment. Unless indicated, all associations are statistically significant at an FDR-corrected $P < 0.05$. "NS" indicates no statistical significance. **C)** Correlation of gbM, unM, and teM genes with chromatin environment. The number of genes and number of TEs were calculated in 100-kb sliding windows with a 50-kb step size and used as a proxy for regions of euchromatin and heterochromatin. Unless indicated, all associations are statistically significant at an FDR-corrected $P < 0.05$. "NS" indicates no statistical significance.

and is characteristic of what we termed "core:multicopy" genes. The latter "duplication-resistant" genes we have termed "core:single-copy" are thought to be under selective pressure to maintain singleton status (Paterson et al. 2006; De Smet et al. 2013; Li et al. 2016). gbM and unM mark different functional sets of core:single-copy genes and likely reflect differences in expression. While core:single-copy genes are predominantly singletons across angiosperms, in each species, some duplicate copies still persist (SC-intermediates). By contrasting these SC-intermediates to core:multicopy genes, we found that SC-intermediates have more frequent differences in genic methylation between duplicate pairs and a higher frequency of teM compared with core:multicopy genes and often duplicate pairs as a whole. SC-intermediates are predominantly syntenic genes resulting from WGD and do

not differ in evolutionary age from core:multicopy genes. As such, they are unlikely to have been silenced prior to WGD, and the gain of teM would not have occurred via movement to heterochromatic regions. Despite being from conserved gene families, SC-intermediates have higher K_a/K_s ratios indicating relaxed selection and are associated with PAV. We propose that SC-intermediates are in the process of fractionation and that silencing by DNA methylation has a role in maintaining dosage and is a first step to their fractionation. However, it is unclear how these conserved genes would become silenced. Experimental approaches, including use of resynthesized or synthetic polyploids (Edger et al. 2017), may be necessary to capture this process.

It has been proposed that silencing by DNA methylation can result in retention of paralogs and their functional divergence

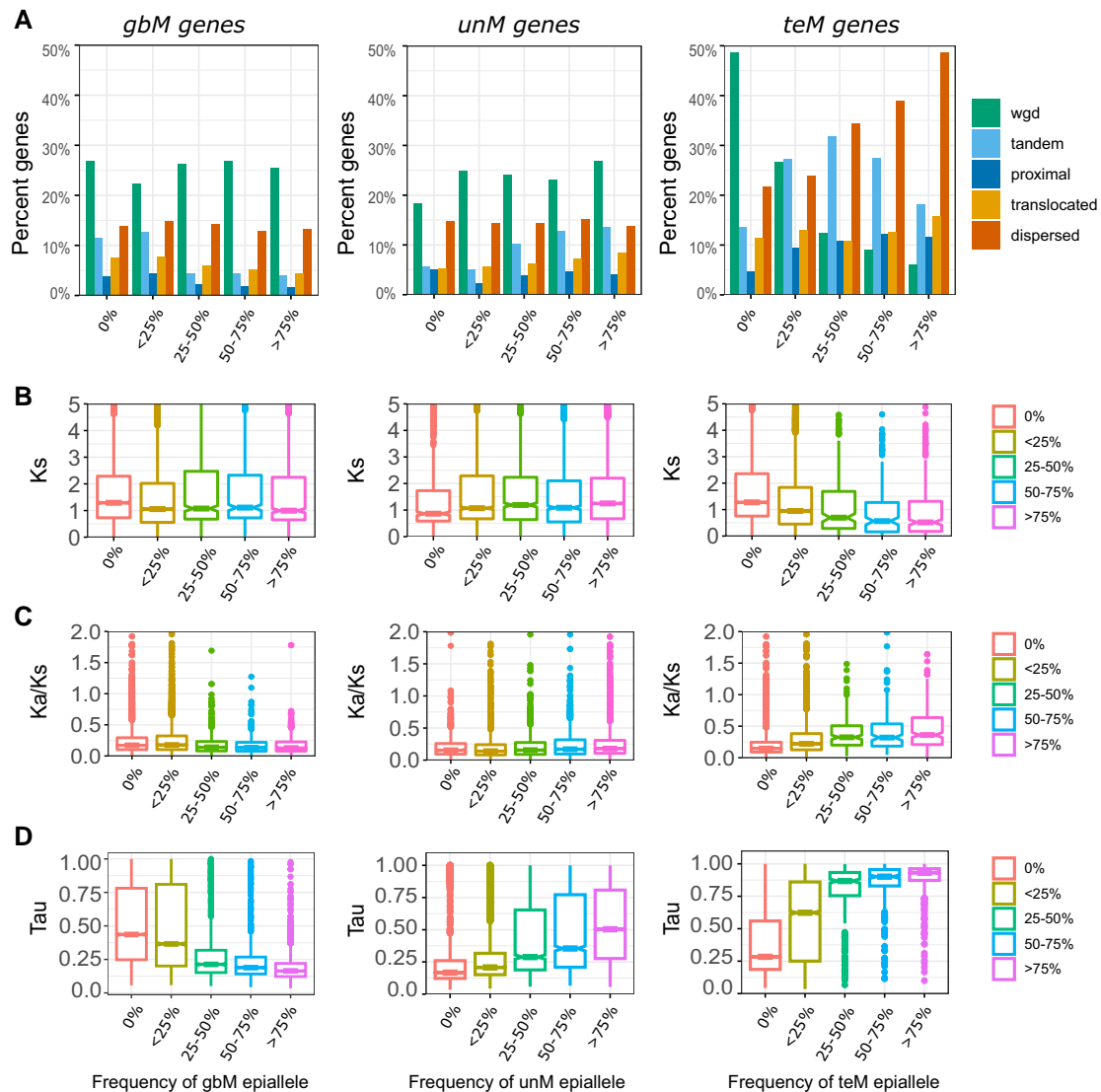


Figure 6. Genic methylation frequency within a population is associated with duplication type, age, sequence evolution, and expression specificity. **A**) The proportion of each type of duplication at different population frequencies of *gbM*, *unM*, and *teM* (0%, <25%, 25% to 50%, 50% to 75%, and >75%) in 928 *A. thaliana* accessions. **B**, **C**, **D**) The distribution of **B**) *Ks* (synonymous substitutions), **C**) *Ka/Ks* (ratio of *Ka*, nonsynonymous substitutions to *Ks*, synonymous substitutions), **D**) and tissue specificity index (τ) for genes at different population frequencies of *gbM*, *unM*, and *teM*. Center line in the boxplot represents the median *Ks* value and *Ka/Ks* ratio, while the box limits represent 25% and 75% percentile of the interquartile range; whiskers represent 1.5 times above or below the interquartile range and dots represent outliers.

(e.g. epigenetic complementation) (Adams et al. 2003; Rodin and Riggs 2003; Chang and Liao 2012). Alternatively, silencing may lead to pseudogenization and gene loss (Hua et al. 2013; El Baidouri et al. 2018). Neither hypothesis is necessarily wrong and cases of both likely exist within a genome. Our results suggest that pseudogenization and loss are the predominant consequences. This is most evident for SGDs and core: single-copy angiosperm genes. However, there is suggestive evidence for epigenetic complementation in the divergence of expression states, and many *teM* genes are still expressed under limited conditions. Epigenetic complementation could also occur through silencing by other chromatin marks, like H3K27me3. Furthermore, many *teM*-containing duplicates have a *Ka/Ks* > 1, possibly indicating

positive selection. Rapid functional divergence of SGDs was observed in grasses; many of these have characteristics similar to *teM* SGDs (Jiang and Assis 2019), but this will require further analysis. Our data show the genic DNA methylation marks differing evolutionary fates of duplicate genes and may have a role in maintaining dosage following gene duplication.

Materials and methods

Genome and methylome data

We used genomes and annotations for 58 angiosperm species (Tuskan et al. 2006; Jaillon et al. 2007; Ming et al. 2008;

Sato et al. 2008; The International Brachypodium Initiative 2010; Schmutz et al. 2010; Hu et al. 2011; Bennetzen et al. 2012; D'Hont et al. 2012; Garcia-Mas et al. 2012; Lamesch et al. 2012; Paterson et al. 2012; Amborella Genome Project 2013; Guo et al. 2013; Hellsten et al. 2013; Kawahara et al. 2013; Ming et al. 2013; Motamayor et al. 2013; Sharma et al. 2013; Singh et al. 2013; Slotte et al. 2013; Yang et al. 2013; Dohm et al. 2014; Liu et al. 2014; Parkin et al. 2014; Schmutz et al. 2014; Tang et al. 2014; Wang, Haberer, et al. 2014; Bartholomé et al. 2015; VanBuren et al. 2015; Bertoli et al. 2016; Bombarely et al. 2016; Bredeson et al. 2016; Cheng et al. 2017; Daccord et al. 2017; Harkess et al. 2017; Jiao et al. 2017; Verde et al. 2017; Xu et al. 2017; Edger et al. 2018; Filiault et al. 2018; Hibrand Saint-Oyant et al. 2018; Hulse-Kemp et al. 2018; Lovell et al. 2018; McCormick et al. 2018; VanBuren et al. 2018; Wu et al. 2018; Xue et al. 2018; Barchi et al. 2019; Colle et al. 2019; Edger et al. 2019; Li et al. 2019; Valliyodan et al. 2019; Hosmani et al. 2019; Mamidi et al. 2019; Lovell et al. 2021), including 43 species (Supplemental Table S1) with whole-genome bisulfite sequencing (WGBS) data (Amborella Genome Project 2013; Seymour et al. 2014; Kim et al. 2015; Ong-Abdullah et al. 2015; Secco et al. 2015; Bertoli et al. 2016; Bewick et al. 2016; Niederhuth et al. 2016; Daccord et al. 2017; Dong et al. 2017; Picard and Gehring 2017; Song et al. 2017; Turco et al. 2017; Lü et al. 2018; Wang et al. 2018; Cheng et al. 2018b; Noshay et al. 2019; Yang et al. 2019) and additional 15 species included as outgroups (Supplemental Table S1). Genes were filtered to remove putative misannotated TEs as previously described (Bowman et al. 2017) with slight modifications. First, genes were searched against Pfam-A using hmmscan (Potter et al. 2018) filtering genes matching a curated list of TE domains (https://github.com/Childs-Lab/GC_specific_MAKER) with an e -value $< 1e-5$. Next, genes were searched against a set of transposase sequences (www.hrt.msu.edu/uploads/535/78637/Tpases020812.gz) using DIAMOND blastp (Buchfink et al. 2015) and hits with an e -value $< 1e-10$ removed.

DNA methylation analyses

WGBS from 43 angiosperm species (Supplemental Table S1) was mapped to their respective genomes using methylpy v1.2.9 (Schultz et al. 2015). Genes were classified as gbM, teM, and unM as previously done with slight modification (Takuno and Gaut 2012; Niederhuth et al. 2016). First, a background rate was calculated for CG, CHG, CHH, and non-CG (combined CHG and CHH) methylation by averaging the percentage of methylated sites in that context across primary transcript coding regions (CDS feature) of all species. Each gene was tested for enrichment of CG, CHG, CHH, or non-CG in its primary transcript coding region against this background rate using a binomial test and P -values corrected for FDR by the Benjamini–Hochberg (BH) procedure (Benjamini and Hochberg 1995). Genes enriched for CG methylation with ≥ 10 CG sites and nonsignificant 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 et al. 2012) $\leq 2\%$ for all contexts (CG, CHG, or CHH) were classified as unM. Genes lacking DNA methylation data were classified “missing” and those with intermediate DNA methylation levels not fitting the above criteria “unclassified.”

Gene duplication classification

Each species was blasted against itself and Amborella (*A. trichopoda*) (outgroup) using double index alignment of next-generation sequencing data (DIAMOND) blastp (Buchfink et al. 2015). *A. thaliana* was used as the outgroup for *A. trichopoda*. Hits from the same orthogroup and an e -value cutoff $< 1e-5$ were retained. Paralogs were classified by DupGen_finder-unique (Qiao et al. 2019), requiring ≥ 5 genes for collinearity and ≤ 10 intervening genes to classify as “proximal” duplicates. MCSanX-transposed (Wang, Li, and Paterson 2013) was used to detect translocated duplicates at different epochs since species divergence (Supplemental Table S12). Genic methylation enrichment in duplication types was determined by a 2-sided Fisher's exact test (Fisher 1934) with FDR correction by BH and plotted using heatmap.2 in gplots. The phylogenetic tree was generated using “V.PhyloMaker” (Jin and Qian 2019) and “phytools” (Revell 2012). To avoid overcounting, if a gene had more than 1 potential paralog, we retained the pair with the lowest e -value.

Orthogroup analyses

Orthogroups were identified for protein sequences of 58 angiosperm species (Supplemental Table S1) using Orthofinder v2.5.2 (Emms and Kelly 2015; Emms and Kelly 2019), with the options “-M dendroblast -S diamond_ultra_sens, -l 1.3.” Orthogroups represented in ≥ 51 species ($\sim 87.9\%$; Supplemental Fig. S2) were classified as “core angiosperm” orthogroups. This accounts for missing annotations and is equivalent to cutoffs in past work (Li et al. 2016). Following Li et al., we classified core angiosperm orthogroups as core:single-copy if represented by a single gene in $\geq 70\%$ species and the remainder as core:multicopy. Remaining orthogroups were classified based on increasing lineage specificity: “cross family” if present in multiple plant families, “family specific” if found in a single plant family, or “species/lineage specific” if limited to a single species. Within each species, a subset of core:single-copy orthogroups still retained duplicate copies and were classified as SC-intermediates and those represented by only a single gene as SC-singletons. A 2-proportion Z-test was used to test differences in genic methylation between SC-intermediates and SC-singletons.

Sequence evolution

The calculate_Ka_Ks_pipeline.pl (Qiao et al. 2019) was used to determine nonsynonymous (K_a) and synonymous substitutions (K_s) for duplicate pairs. Protein sequences are aligned

by MAFFT (v7.402) (Kato and Standley 2013), converted to a codon alignment with PAL2NAL (Suyama et al. 2006), and KaKs_Calculator 2.0 used to calculate Ka, Ks, and Ka/Ks with the γ -MYN method (Wang et al. 2010; Qiao et al. 2019). 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). 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 2-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 expression atlases (O'Rourke et al. 2014; Klepikova et al. 2016; McCormick et al. 2018; Wang et al. 2019). *A. thaliana* reads were downloaded from NCBI SRA (PRJNA314076 and PRJNA324514), mapped with STAR (Dobin et al. 2013), and normalized for library size in DESeq2 (Love et al. 2014). *G. max*, *P. vulgaris*, and *S. bicolor* normalized data were downloaded from Phytozome (Goodstein et al. 2012). The tissue specificity index (τ) was calculated in R for each gene as previously described (Yanai et al. 2005). Genes not expressed under any conditions were excluded as τ could not be calculated. Pearson correlation coefficients were calculated for each duplicate pair in R.

Transposons and genomic distribution

TEs were annotated de novo for all species using extensive de novo TE annotator (EDTA) (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).

Arabidopsis diversity

WGBS data for 928 *A. thaliana* accessions, previously aligned by methylpy (Kawakatsu et al. 2016), were downloaded from the Gene Expression Omnibus (GEO Accession GSE43857). Genes were classified as before and the frequency of gbM/unM/teM in the population calculated for each gene.

Accession numbers

NCBI SRA accession numbers for the data sets used in the study are listed in Supplemental Table S1.

Acknowledgments

We thank Dr. Patrick Edger for the unpublished *Cleome violaceae* genome and Dr. Leslie Kollar for critical reading and comments on the manuscript.

Author contributions

S.K.K.R. and C.E.N. designed the work and analyses. 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.

Supplemental data

The following materials are available in the online version of this article.

Supplemental Figure S1. Schematic representation of different orthogroup classifications.

Supplemental Figure S2. Distribution of orthogroups across 58 angiosperm species.

Supplemental Figure S3. Distribution of orthogroups in genic methylation classes.

Supplemental Figure S4. Enrichment or depletion of different genic methylation classes (gbM, unM, and teM) for each orthogroup category (core:multicopy, core:single-copy, cross-family, family-specific, and lineage/species-specific).

Supplemental Figure S5. Enrichment or depletion of different types of duplicates (whole-genome and different types of SGDs) for each orthogroup category (core:multicopy, core:single-copy, cross-family, family-specific, and lineage/species-specific).

Supplemental Figure S6. Proportion of paralogs with similar and divergent DNA methylation profiles.

Supplemental Figure S7. Distribution of genic methylation classified genes based on synonymous substitution (Ks) across different types of gene duplicate pairs.

Supplemental Figure S8. Proportion of different genic methylation in transposed duplicates across different epochs.

Supplemental Figure S9. Distribution of core:multicopy and core:single-copy (intermediate) paralogs based on synonymous substitutions (Ks).

Supplemental Figure S10. 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.

Supplemental Figure S11. Distribution of core:multicopy and core:single-copy (intermediate) paralogs based on the ratio of nonsynonymous substitution (Ka), with synonymous substitutions (Ks).

Supplemental Figure S12. Percentage of total (all genes), gbM, teM, and unM genes with known PAV.

Supplemental Figure S13. Proportion of SC-singletons and SC-intermediates in PAV genes in *B. oleracea*, *S. lycopersicum*, *S. tuberosum*, and *Z. mays*.

Supplemental Figure S14. Tau specificity of gbM, teM, and unM genes in *G. max*, *P. vulgaris*, and *S. bicolor*.

Supplemental Figure S15. Tau specificity of gbM, teM, and unM genes for each orthogroup category in *A. thaliana*, *G. max*, *P. vulgaris*, and *S. bicolor*.

Supplemental Figure S16. Tau specificities of different types of duplicate genes in *A. thaliana*, *G. max*, *P. vulgaris*, and *S. bicolor*.

Supplemental Figure S17. Half plots showing 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*.

Supplemental Figure S18. Absolute differences in tau specificity between duplicate pairs in *G. max*, *P. vulgaris*, and *S. bicolor*.

Supplemental Figure S19. 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*.

Supplemental Figure S20. Proportion of different orthogroup classifications and gbM/unM/teM epiallele frequency within *A. thaliana* population.

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

Supplemental Table S2. Classification of genic methylation of all genes in each species.

Supplemental Table S3. Number of genes derived from different types of duplications in each species.

Supplemental Table S4. Enrichment and depletion of genic methylation classifications across different types of gene duplicates.

Supplemental Table S5. Enrichment and depletion of genic methylation classified genes across different orthogroup classifications.

Supplemental Table S6. Enrichment and depletion of different types of gene duplicates across each orthogroup category.

Supplemental Table S7. Number of duplicate gene pairs with different or the same genic methylation status for each type of duplication in each species.

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

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

Supplemental Table S10. Top 5 functional categories for each GO enrichment comparison for gbM core:single-copy genes and unM core:single-copy genes across 33 species.

Supplemental Table S11. Duplicate pair classifications ranked based on median Ks values for SGDs and WGDs.

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

Supplemental Table S13. Enrichment and depletion of genic methylation classifications across different epochs of transposed duplicates for all species.

Supplemental Table S14. Duplicate pair classifications ranked based on median Ka/Ks values for SGDs and WGDs.

Supplemental Table S15. Enrichment and depletion of different types of duplicate genes with Ka/Ks ratio > 1.0.

Supplemental Table S16. Enrichment and depletion of different classifications of duplicate gene pairs with Ka/Ks ratio > 1.0.

Supplemental Table S17. Enrichment and depletion of known presence–absence variants for gbM, teM, and unM genes.

Supplemental Table S18. Differences in the distribution of absolute difference in tau for duplicate gene pairs with similar or divergence in methylation.

Supplemental Table S19. Enrichment and depletion of TEs with gbM, teM, and unM paralogs in each species.

Supplemental Table S20. Enrichment and depletion of TEs with different types of duplication in each species.

Supplemental Table S21. Presence/absence of TEs in the gene body and 1-kb upstream and 1-kb downstream for duplicate paralogs differing in their genic methylation.

Supplemental Table S22. Correlations between genic methylation classes (gbM, teM, and unM) and genomic features (number of genes, TEs, and TE base pairs) in 100-kb sliding windows with a 50-kb step size.

Funding

This work was supported by Michigan State University, the USDA National Institute of Food and Agriculture (MICL02572), and the National Science Foundation (IOS-2029959). M.L. was supported by the National Science Foundation (DBI-1757043).

Conflict of interest statement. None declared.

Data availability

Raw data sources are listed in [Supplemental Table S1](#). Formatted genomes and data are available at DataDryad <https://doi.org/10.5061/dryad.n8pk0p30v>. Code and scripts used in these analyses are available at: <https://github.com/niederhuth/DNA-methylation-signatures-of-duplicate-gene-evolution-in-angiosperms>.

References

- Adams KL, Cronn R, Percifield R, Wendel JF. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc Natl Acad Sci U S A*. 2003;100(8):4649–4654. <https://doi.org/10.1073/pnas.0630618100>
- Amborella Genome Project. The *Amborella* genome and the evolution of flowering plants. *Science* 2013;342(6165):1241089. <https://doi.org/10.1126/science.1241089>
- Barchi L, Pietrella M, Venturini L, Minio A, Toppino L, Acquadro A, Andolfo G, Aprea G, Avanzato C, Bassolino L, et al. A chromosome-anchored eggplant genome sequence reveals key events in Solanaceae evolution. *Sci Rep*. 2019;9(1):11769. <https://doi.org/10.1038/s41598-019-47985-w>
- Bartholomé J, Mandrou E, Mabiala A, Jenkins J, Nabihoudine I, Klopp C, Schmutz J, Plomion C, Gion J-M. High-resolution genetic maps of *Eucalyptus* improve *Eucalyptus grandis* genome assembly. *New Phytol*. 2015;206(4):1283–1296. <https://doi.org/10.1111/nph.13150>
- Becker C, Weigel D. Epigenetic variation: origin and transgenerational inheritance. *Curr Opin Plant Biol*. 2012;15(5):562–567. <https://doi.org/10.1016/j.pbi.2012.08.004>
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Methodol*. 1995;57(1):289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>

- Bennetzen JL, Schmutz J, Wang H, Percifield R, Hawkins J, Pontaroli AC, Estep M, Feng L, Vaughn JN, Grimwood J, et al.** Reference genome sequence of the model plant *Setaria*. *Nat Biotechnol*. 2012;**30**(6):555–561. <https://doi.org/10.1038/nbt.2196>
- Bertioli DJ, Cannon SB, Froenicke L, Huang G, Farmer AD, Cannon EKS, Liu X, Gao D, Clevenger J, Dash S, et al.** The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nat Genet*. 2016;**48**(4):438–446. <https://doi.org/10.1038/ng.3517>
- Bewick AJ, Ji L, Niederhuth CE, Willing E-M, Hofmeister BT, Shi X, Wang L, Lu Z, Rohr NA, Hartwig B, et al.** On the origin and evolutionary consequences of gene body DNA methylation. *Proc Natl Acad Sci U S A*. 2016;**113**(32):9111–9116. <https://doi.org/10.1073/pnas.1604666113>
- Birchler JA, Veitia RA.** The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. *New Phytol*. 2010;**186**(1):54–62. <https://doi.org/10.1111/j.1469-8137.2009.03087.x>
- Bombarely A, Moser M, Amrad A, Bao M, Bapaume L, Barry CS, Bliker M, Boersma MR, Borghi L, Bruggmann R, et al.** Insight into the evolution of the Solanaceae from the parental genomes of *Petunia hybrida*. *Nat Plants*. 2016;**2**(6):16074. <https://doi.org/10.1038/nplants.2016.74>
- Bowman MJ, Pulman JA, Liu TL, Childs KL.** A modified GC-specific MAKER gene annotation method reveals improved and novel gene predictions of high and low GC content in *Oryza sativa*. *BMC Bioinformatics* 2017;**18**(1):522. <https://doi.org/10.1186/s12859-017-1942-z>
- Bredeson JV, Lyons JB, Prochnik SE, Wu GA, Ha CM, Edsinger-Gonzales E, Grimwood J, Schmutz J, Rabbi IY, Egesi C, et al.** Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nat Biotechnol*. 2016;**34**(5):562–570. <https://doi.org/10.1038/nbt.3535>
- Bridges CB.** Salivary chromosome maps. *Journal of Heredity* 1935;**26**(2):60–64. <https://doi.org/10.1093/oxfordjournals.jhered.a104022>
- Buchfink B, Xie C, Huson DH.** Fast and sensitive protein alignment using DIAMOND. *Nat Methods*. 2015;**12**(1):59–60. <https://doi.org/10.1038/nmeth.3176>
- Chang AY-F, Liao B-Y.** DNA methylation rebalances gene dosage after mammalian gene duplications. *Mol Biol Evol*. 2012;**29**(1):133–144. <https://doi.org/10.1093/molbev/msr174>
- Cheng C-Y, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD.** Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J*. 2017;**89**(4):789–804. <https://doi.org/10.1111/tpj.13415>
- Cheng J, Niu Q, Zhang B, Chen K, Yang R, Zhu J-K, Zhang Y, Lang Z.** Downregulation of RdDM during strawberry fruit ripening. *Genome Biol*. 2018b;**19**(1):212. <https://doi.org/10.1186/s13059-018-1587-x>
- Cheng F, Wu J, Cai X, Liang J, Freeling M, Wang X.** Gene retention, fractionation and subgenome differences in polyploid plants. *Nat Plants*. 2018a;**4**(5):258–268. <https://doi.org/10.1038/s41477-018-0136-7>
- Colle M, Leisner CP, Wai CM, Ou S, Bird KA, Wang J, Wisecaver JH, Yocca AE, Alger EI, Tang H, et al.** Haplotype-phased genome and evolution of phytonutrient pathways of tetraploid blueberry. *Gigascience* 2019;**8**(3):giz012. <https://doi.org/10.1093/gigascience/giz012>
- Conant GC, Birchler JA, Pires JC.** Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Curr Opin Plant Biol*. 2014;**19**:91–98. <https://doi.org/10.1016/j.pbi.2014.05.008>
- Cusack BP, Wolfe KH.** Not born equal: increased rate asymmetry in re-located and retrotransposed rodent gene duplicates. *Mol Biol Evol*. 2006;**24**(3):679–686. <https://doi.org/10.1093/molbev/msl199>
- Daccord N, Celton J-M, Linsmith G, Becker C, Choisne N, Schijlen E, van de Geest H, Bianco L, Micheletti D, Velasco R, et al.** High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nat Genet*. 2017;**49**(7):1099–1106. <https://doi.org/10.1038/ng.3886>
- De Smet R, Adams KL, Vandepoele K, Van Montagu MCE, Maere S, Van de Peer Y.** Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc Natl Acad Sci U S A*. 2013;**110**(8):2898–2903. <https://doi.org/10.1073/pnas.1300127110>
- D’Hont A, Denoed F, Aury J-M, Baurens F-C, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M, et al.** The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* 2012;**488**(7410):213–217. <https://doi.org/10.1038/nature11241>
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR.** STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**(1):15–21. <https://doi.org/10.1093/bioinformatics/bts635>
- Dohm JC, Minoche AE, Holtgräwe D, Capella-Gutiérrez S, Zakrzewski F, Tafer H, Rupp O, Sørensen TR, Stracke R, Reinhardt R, et al.** The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). *Nature* 2014;**505**(7484):546–549. <https://doi.org/10.1038/nature12817>
- Dong P, Tu X, Chu P-Y, Lü P, Zhu N, Grierson D, Du B, Li P, Zhong S.** 3D chromatin architecture of large plant genomes determined by local A/B compartments. *Mol Plant*. 2017;**10**(12):1497–1509. <https://doi.org/10.1016/j.molp.2017.11.005>
- Edger PP, Poorten TJ, VanBuren R, Hardigan MA, Colle M, McKain MR, Smith RD, Teresi SJ, Nelson ADL, Wai CM, et al.** Origin and evolution of the octoploid strawberry genome. *Nat Genet*. 2019;**51**(3):541–547. <https://doi.org/10.1038/s41588-019-0356-4>
- Edger PP, Smith R, McKain MR, Cooley AM, Vallejo-Marin M, Yuan Y, Bewick AJ, Ji L, Platts AE, Bowman MJ, et al.** Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140-year-old naturally established neo-allopolyploid monkeyflower. *Plant Cell* 2017;**29**(9):2150–2167. <https://doi.org/10.1105/tpc.17.00010>
- Edger PP, VanBuren R, Colle M, Poorten TJ, Wai CM, Niederhuth CE, Alger EI, Ou S, Acharya CB, Wang J, et al.** Single-molecule sequencing and optical mapping yields an improved genome of woodland strawberry (*Fragaria vesca*) with chromosome-scale contiguity. *Gigascience* 2018;**7**(2):1–7. <https://doi.org/10.1093/gigascience/gix124>
- Eichten SR, Ellis NA, Makarevitch I, Yeh C-T, Gent JI, Guo L, McGinnis KM, Zhang X, Schnable PS, Vaughn MW, et al.** Spreading of heterochromatin is limited to specific families of maize retrotransposons. *PLoS Genet*. 2012;**8**(12):e1003127. <https://doi.org/10.1371/journal.pgen.1003127>
- El Baidouri M, Kim KD, Abernathy B, Li Y-H, Qiu L-J, Jackson SA.** Genic C-methylation in soybean is associated with gene paralogs re-located to transposable element-rich pericentromeres. *Mol Plant*. 2018;**11**(3):485–495. <https://doi.org/10.1016/j.molp.2018.02.006>
- Emms DM, Kelly S.** Orthofinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol*. 2015;**16**(1):157. <https://doi.org/10.1186/s13059-015-0721-2>
- Emms DM, Kelly S.** Orthofinder: phylogenetic orthology inference for comparative genomics. *Genome Biol*. 2019;**20**(1):238. <https://doi.org/10.1186/s13059-019-1832-y>
- Feng S, Cokus SJ, Zhang X, Chen P-Y, Bostick M, Goll MG, Hetzel J, Jain J, Strauss SH, Halpern ME, et al.** Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci U S A*. 2010;**107**(19):8689–8694. <https://doi.org/10.1073/pnas.1002720107>
- Ferris SD, Whitt GS.** Evolution of the differential regulation of duplicate genes after polyploidization. *J Mol Evol*. 1979;**12**(4):267–317. <https://doi.org/10.1007/BF01732026>
- Filialt DL, Ballerini ES, Mandáková T, Aköz G, Derieg NJ, Schmutz J, Jenkins J, Grimwood J, Shu S, Hayes RD, et al.** The genome provides insight into adaptive radiation and reveals an extraordinarily

- polymorphic chromosome with a unique history. *eLife* 2018(7): e36426. <https://doi.org/10.7554/eLife.36426>
- Fisher SRA.** Statistical methods for research workers. 1934.
- Flagel LE, Wendel JF.** Gene duplication and evolutionary novelty in plants. *New Phytologist*. 2009;183(3):557–564. <https://doi.org/10.1111/j.1469-8137.2009.02923.x>
- Freeling M.** Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annu Rev Plant Biol*. 2009;60(1):433–453. <https://doi.org/10.1146/annurev.arplant.043008.092122>
- Freeling M, Lyons E, Pedersen B, Alam M, Ming R, Lisch D.** Many or most genes in *Arabidopsis* transposed after the origin of the order Brassicales. *Genome Res*. 2008;18(12):1924–1937. <https://doi.org/10.1101/gr.081026.108>
- Ganko EW, Meyers BC, Vision TJ.** Divergence in expression between duplicated genes in *Arabidopsis*. *Mol Biol Evol*. 2007;24(10): 2298–2309. <https://doi.org/10.1093/molbev/msm158>
- Gao L, Gonda I, Sun H, Ma Q, Bao K, Tieman DM, Burzynski-Chang EA, Fish TL, Stromberg KA, Sacks GL, et al.** The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat Genet*. 2019;51(6):1044–1051. <https://doi.org/10.1038/s41588-019-0410-2>
- Garcia-Mas J, Benjak A, Sanseverino W, Bourgeois M, Mir G, González VM, Hénaff E, Câmara F, Cozzuto L, Lowy E, et al.** The genome of melon (*Cucumis melo* L.). *Proc Natl Acad Sci U S A*. 2012;109(29):11872–11877. <https://doi.org/10.1073/pnas.1205415109>
- Golicz AA, Bayer PE, Barker GC, Edger PP, Kim H, Martinez PA, Chan CKK, Severn-Ellis A, McCombie WR, Parkin IAP, et al.** The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nat Commun*. 2016;7(1):13390. <https://doi.org/10.1038/ncomms13390>
- Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, et al.** Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res*. 2012;40(D1):D1178–D1186. <https://doi.org/10.1093/nar/gkr944>
- Guo S, Zhang J, Sun H, Salse J, Lucas WJ, Zhang H, Zheng Y, Mao L, Ren Y, Wang Z, et al.** The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. *Nat Genet*. 2013;45(1):51–58. <https://doi.org/10.1038/ng.2470>
- Hardigan MA, Crisovan E, Hamilton JP, Kim J, Laimbeer P, Leisner CP, Manrique-Carpintero NC, Newton L, Pham GM, Vaillancourt B, et al.** Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated *Solanum tuberosum*. *Plant Cell* 2016;28(2):388–405. <https://doi.org/10.1105/tpc.15.00538>
- Harkess A, Zhou J, Xu C, Bowers JE, Van der Hulst R, Ayyampalayam S, Mercati F, Riccardi P, McKain MR, Kakrana A, et al.** The asparagus genome sheds light on the origin and evolution of a young Y chromosome. *Nat Commun*. 2017;8(1):1279. <https://doi.org/10.1038/s41467-017-01064-8>
- Hellsten U, Wright KM, Jenkins J, Shu S, Yuan Y, Wessler SR, Schmutz J, Willis JH, Rokhsar DS.** Fine-scale variation in meiotic recombination in *Mimulus* inferred from population shotgun sequencing. *Proc Natl Acad Sci U S A*. 2013;110(48):19478–19482. <https://doi.org/10.1073/pnas.1319032110>
- Hibrand Saint-Oyant L, Ruttink T, Hamama L, Kirov I, Lakhwani D, Zhou NN, Bourke PM, Daccord N, Leus L, Schulz D, et al.** A high-quality genome sequence of *Rosa chinensis* to elucidate ornamental traits. *Nat Plants*. 2018;4(7):473–484. <https://doi.org/10.1038/s41477-018-0166-1>
- Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, Peñagaricano F, Lindquist E, Pedraza MA, Barry K, et al.** Insights into the maize pan-genome and pan-transcriptome. *Plant Cell* 2014;26(1):121–135. <https://doi.org/10.1105/tpc.113.119982>
- Hirsch CD, Springer NM.** Transposable element influences on gene expression in plants. *Biochim Biophys Acta Gene Regul Mech*. 2017;1860(1):157–165. <https://doi.org/10.1016/j.bbagr.2016.05.010>
- Hosmani PS, Flores-Gonzalez M, van de Geest H, Maumus F, Bakker LV, Schijlen E, van Haarst J, Cordewener J, Sanchez-Perez G, Peters S, et al.** An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-C proximity ligation and optical maps. *bioRxiv* 2019:767764. <https://doi.org/10.1101/767764>
- Hu TT, Pattyn P, Bakker EG, Cao J, Cheng J-F, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, et al.** The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat Genet*. 2011;43(5):476–481. <https://doi.org/10.1038/ng.807>
- Hua Z, Pool JE, Schmitz RJ, Schultz MD, Shiu S-H, Ecker JR, Vierstra RD.** Epigenomic programming contributes to the genomic drift evolution of the F-Box protein superfamily in *Arabidopsis*. *Proc Natl Acad Sci U S A*. 2013;110(42):16927–16932. <https://doi.org/10.1073/pnas.1316009110>
- Hulse-Kemp AM, Maheshwari S, Stoffel K, Hill TA, Jaffe D, Williams SR, Weisenfeld N, Ramakrishnan S, Kumar V, Shah P, et al.** Reference quality assembly of the 3.5-Gb genome of *Capsicum annuum* from a single linked-read library. *Hortic Res*. 2018;5(4). <https://doi.org/10.1038/s41438-017-0011-0>
- Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Casagrande A, Choise N, Aubourg S, Vitulo N, Jubin C, et al.** The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 2007;449(7161):463–467. <https://doi.org/10.1038/nature06148>
- Jiang X, Assis R.** Rapid functional divergence after small-scale gene duplication in grasses. *BMC Evol Biol*. 2019;19(1):97. <https://doi.org/10.1186/s12862-019-1415-2>
- Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, Campbell MS, Stein JC, Wei X, Chin C-S, et al.** Improved maize reference genome with single-molecule technologies. *Nature* 2017;546(7659):524–527. <https://doi.org/10.1038/nature22971>
- Jin Y, Qian H.** VPhyloMaker: an R package that can generate very large phylogenies for vascular plants. *Ecography* 2019;42(8):1353–1359. <https://doi.org/10.1111/ecog.04434>
- Katoh K, Standley DM.** MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013;30(4):772–780. <https://doi.org/10.1093/molbev/mst010>
- Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, Schwartz DC, Tanaka T, Wu J, Zhou S, et al.** Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice* 2013;6(1):4. <https://doi.org/10.1186/1939-8433-6-4>
- Kawakatsu T, Huang S-SC, Jupe F, Sasaki E, Schmitz RJ, Urlich MA, Castanon R, Nery JR, Barragan C, He Y, et al.** Epigenomic diversity in a global collection of *Arabidopsis thaliana* accessions. *Cell* 2016;166(2):492–505. <https://doi.org/10.1016/j.cell.2016.06.044>
- Keller TE, Yi SV.** DNA methylation and evolution of duplicate genes. *Proc Natl Acad Sci U S A*. 2014;111(16):5932–5937. <https://doi.org/10.1073/pnas.1321420111>
- Kim KD, El Baidouri M, Abernathy B, Iwata-Otsubo A, Chavarro C, Gonzales M, Libault M, Grimwood J, Jackson SA.** A comparative epigenomic analysis of polyploidy-derived genes in soybean and common bean. *Plant Physiol*. 2015;168(4):1433–1447. <https://doi.org/10.1104/pp.15.00408>
- Klepkova AV, Kasianov AS, Gerasimov ES, Logacheva MD, Penin AA.** A high resolution map of the *Arabidopsis thaliana* developmental transcriptome based on RNA-seq profiling. *Plant J*. 2016;88(6): 1058–1070. <https://doi.org/10.1111/tpj.13312>
- Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, et al.** The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res*. 2012;40(D1):D1202–D1210. <https://doi.org/10.1093/nar/gkr1090>
- Li Q, Li H, Huang W, Xu Y, Zhou Q, Wang S, Ruan J, Huang S, Zhang Z.** A chromosome-scale genome assembly of cucumber (*Cucumis*

- sativus* L.). *Gigascience* 2019;8(6):giz072. <https://doi.org/10.1093/gigascience/giz072>
- Li W-H, Yang J, Gu X. Expression divergence between duplicate genes. *Trends Genet.* 2005;21(11):602–607. <https://doi.org/10.1016/j.tig.2005.08.006>
- Li Z, Defoort J, Tasdighian S, Maere S, Van de Peer Y, De Smet R. Gene duplicability of core genes is highly consistent across all angiosperms. *Plant Cell* 2016;28(2):326–344. <https://doi.org/10.1105/tpc.15.00877>
- Liu M-J, Zhao J, Cai Q-L, Liu G-C, Wang J-R, Zhao Z-H, Liu P, Dai L, Yan G, Wang W-J, et al. The complex jujube genome provides insights into fruit tree biology. *Nat Commun.* 2014;5(1):5315. <https://doi.org/10.1038/ncomms6315>
- Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15(12):550. <https://doi.org/10.1186/s13059-014-0550-8>
- Lovell JT, Jenkins J, Lowry DB, Mamidi S, Sreedasyam A, Weng X, Barry K, Bonnette J, Campitelli B, Daum C, et al. The genomic landscape of molecular responses to natural drought stress in *Panicum hallii*. *Nat Commun.* 2018;9(1):5213. <https://doi.org/10.1038/s41467-018-07669-x>
- Lovell JT, MacQueen AH, Mamidi S, Bonnette J, Jenkins J, Napier JD, Sreedasyam A, Healey A, Session A, Shu S, et al. Genomic mechanisms of climate adaptation in polyploid bioenergy switchgrass. *Nature* 2021;590(7846):438–444. <https://doi.org/10.1038/s41586-020-03127-1>
- Lü P, Yu S, Zhu N, Chen Y-R, Zhou B, Pan Y, Tzeng D, Fabi JP, Argyris J, Garcia-Mas J, et al. Genome encode analyses reveal the basis of convergent evolution of fleshy fruit ripening. *Nat Plants.* 2018;4(10):784–791. <https://doi.org/10.1038/s41477-018-0249-z>
- Lynch M. Genomics. Gene duplication and evolution. *Science* 2002;297(5583):945–947. <https://doi.org/10.1126/science.1075472>
- Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. *Science* 2000;290(5494):1151–1155. <https://doi.org/10.1126/science.290.5494.1151>
- Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y. Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci U S A.* 2005;102(15):5454–5459. <https://doi.org/10.1073/pnas.0501102102>
- Mamidi S, Healey A, Huang P, Grimwood J, Jenkins J, Barry K, Sreedasyam A, Shu S, Lovell JT, Feldman M, et al. The *Setaria viridis* genome and diversity panel enables discovery of a novel domestication gene. *bioRxiv* 2019:744557. <https://doi.org/10.1101/744557>
- McCormick RF, Truong SK, Sreedasyam A, Jenkins J, Shu S, Sims D, Kennedy M, Amirebrahimi M, Weers BD, McKinley B, et al. The *Sorghum bicolor* reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *Plant J.* 2018;93(2):338–354. <https://doi.org/10.1111/tpj.13781>
- Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KLT, et al. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* 2008;452(7190):991–996. <https://doi.org/10.1038/nature06856>
- Ming R, VanBuren R, Liu Y, Yang M, Han Y, Li L-T, Zhang Q, Kim M-J, Schatz MC, Campbell M, et al. Genome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.). *Genome Biol.* 2013;14(5):R41. <https://doi.org/10.1186/gb-2013-14-5-r41>
- Miyata T, Yasunaga T. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *J Mol Evol.* 1980;16(1):23–36. <https://doi.org/10.1007/BF01732067>
- Motamayor JC, Mockaitis K, Schmutz J, Haiminen N, Livingstone D III, Cornejo O, Findley SD, Zheng P, Utro F, Royaert S, et al. The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. *Genome Biol.* 2013;14(6):r53. <https://doi.org/10.1186/gb-2013-14-6-r53>
- Niederhuth CE, Bewick AJ, Ji L, Alabady MS, Kim KD, Li Q, Rohr NA, Rambani A, Burke JM, Udall JA, et al. Widespread natural variation of DNA methylation within angiosperms. *Genome Biol.* 2016;17(1):194. <https://doi.org/10.1186/s13059-016-1059-0>
- Niederhuth CE, Schmitz RJ. Putting DNA methylation in context: from genomes to gene expression in plants. *Biochim Biophys Acta Gene Regul Mech.* 2017;1860(1):149–156. <https://doi.org/10.1016/j.bbagr.2016.08.009>
- Noshay JM, Anderson SN, Zhou P, Ji L, Ricci W, Lu Z, Stitzer MC, Crisp PA, Hirsch CN, Zhang X, et al. Monitoring the interplay between transposable element families and DNA methylation in maize. *PLoS Genet.* 2019;15(9):e1008291. <https://doi.org/10.1371/journal.pgen.1008291>
- Ohno S. Evolution by gene duplication. Berlin, Heidelberg: Springer; 1970. <https://doi.org/10.1007/978-3-642-86659-3>
- Ong-Abdullah M, Ordway JM, Jiang N, Ooi S-E, Kok S-Y, Sarpan N, Azimi N, Hashim AT, Ishak Z, Rosli SK, et al. Loss of *Karma* transposon methylation underlies the mantled somaclonal variant of oil palm. *Nature* 2015;525(7570):533–537. <https://doi.org/10.1038/nature15365>
- O'Rourke JA, Iniguez LP, Fu F, Bucciarelli B, Miller SS, Jackson SA, McClean PE, Li J, Dai X, Zhao PX, et al. An RNA-Seq based gene expression atlas of the common bean. *BMC Genomics* 2014;15(1):866. <https://doi.org/10.1186/1471-2164-15-866>
- Otto SP, Whitton J. Polyploid incidence and evolution. *Annu Rev Genet.* 2000;34(1):401–437. <https://doi.org/10.1146/annurev.genet.34.1.401>
- Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson T, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* 2019;20(1):275. <https://doi.org/10.1186/s13059-019-1905-y>
- Panchy N, Lehti-Shiu MD, Shiu S-H. Evolution of gene duplication in plants. *Plant Physiol.* 2016;171(4):2294–2316. <https://doi.org/10.1104/pp.16.00523>
- Parkin IAP, Koh C, Tang H, Robinson SJ, Kagale S, Clarke WE, Town CD, Nixon J, Krishnakumar V, Bidwell SL, et al. Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid *Brassica oleracea*. *Genome Biol.* 2014;15(6):R77. <https://doi.org/10.1186/gb-2014-15-6-r77>
- Paterson AH, Chapman BA, Kissinger JC, Bowers JE, Feltus FA, Estill JC. Many gene and domain families have convergent fates following independent whole-genome duplication events in Arabidopsis, Oryza, Saccharomyces and Tetraodon. *Trends Genet.* 2006;22(11):597–602. <https://doi.org/10.1016/j.tig.2006.09.003>
- Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, Llewellyn D, Showmaker KC, Shu S, Udall J, et al. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature* 2012;492(7429):423–427. <https://doi.org/10.1038/nature11798>
- Picard CL, Gehring M. Proximal methylation features associated with nonrandom changes in gene body methylation. *Genome Biol.* 2017;18(1):73. <https://doi.org/10.1186/s13059-017-1206-2>
- Pophaly SD, Tellier A. Population level purifying selection and gene expression shape subgenome evolution in maize. *Mol Biol Evol.* 2015;32(12):3226–3235. <https://doi.org/10.1093/molbev/msv191>
- Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. HMMER Web server: 2018 update. *Nucleic Acids Res.* 2018;46(W1):W200–W204. <https://doi.org/10.1093/nar/gky448>
- Qiao X, Li Q, Yin H, Qi K, Li L, Wang R, Zhang S, Paterson AH. Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. *Genome Biol.* 2019;20(1):38. <https://doi.org/10.1186/s13059-019-1650-2>
- Raju SKK, Ritter EJ, Niederhuth CE. Establishment, maintenance, and biological roles of non-CG methylation in plants. *Essays Biochem.* 2019;63(6):743–755. <https://doi.org/10.1042/EBC20190032>
- Revell LJ. Phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* 2012;3(2):217–223. <https://doi.org/10.1111/j.2041-210X.2011.00169.x>

- Rodin SN, Riggs AD. Epigenetic silencing may aid evolution by gene duplication. *J Mol Evol.* 2003;**56**(6):718–729. <https://doi.org/10.1007/s00239-002-2446-6>
- Sato S, Nakamura Y, Kaneko T, Asamizu E, Kato T, Nakao M, Sasamoto S, Watanabe A, Ono A, Kawashima K, et al. Genome structure of the legume, *Lotus japonicus*. *DNA Res.* 2008;**15**(4):227–239. <https://doi.org/10.1093/dnares/dsn008>
- Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, et al. Genome sequence of the paleopolyploid soybean. *Nature* 2010;**463**(7278):178–183. <https://doi.org/10.1038/nature08670>
- Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, Jenkins J, Shu S, Song Q, Chavarro C, et al. A reference genome for common bean and genome-wide analysis of dual domestications. *Nat Genet.* 2014;**46**(7):707–713. <https://doi.org/10.1038/ng.3008>
- Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, Rajagopal N, Nery JR, Ulrich MA, Chen H, et al. Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* 2015;**523**(7559):212–216. <https://doi.org/10.1038/nature14465>
- Schultz MD, Schmitz RJ, Ecker JR. “Leveling” the playing field for analyses of single-base resolution DNA methylomes. *Trends Genet.* 2012;**28**(12):583–585. <https://doi.org/10.1016/j.tig.2012.10.012>
- Secco D, Wang C, Shou H, Schultz MD, Chiarenza S, Nussaume L, Ecker JR, Whelan J, Lister R. Stress induced gene expression drives transient DNA methylation changes at adjacent repetitive elements. *eLife* 2015;**4**:e09343. <https://doi.org/10.7554/eLife.09343>
- Seymour DK, Koenig D, Hagmann J, Becker C, Weigel D. Evolution of DNA methylation patterns in the Brassicaceae is driven by differences in genome organization. *PLoS Genet.* 2014;**10**(11):e1004785. <https://doi.org/10.1371/journal.pgen.1004785>
- Sharma SK, Bolser D, de Boer J, Sønderkær M, Amoros W, Carboni MF, D’Ambrosio JM, de la Cruz G, Di Genova A, Douches DS, et al. Construction of reference chromosome-scale pseudomolecules for potato: integrating the potato genome with genetic and physical maps. *G3 Genes|Genomes|Genetics* 2013;**3**(11):2031–2047. <https://doi.org/10.1534/g3.113.007153>
- Singh R, Ong-Abdullah M, Low E-TL, Manaf MAA, Rosli R, Nookiah R, Ooi LC-L, Ooi S-E, Chan K-L, Halim MA, et al. Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds. *Nature* 2013;**500**(7462):335–339. <https://doi.org/10.1038/nature12309>
- Slotte T, Hazzouri KM, Ågren JA, Koenig D, Maumus F, Guo Y-L, Steige K, Platts AE, Escobar JS, Newman LK, et al. The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nat Genet.* 2013;**45**(7):831–835. <https://doi.org/10.1038/ng.2669>
- Soltis PS, Marchant DB, Van de Peer Y, Soltis DE. Polyploidy and genome evolution in plants. *Curr Opin Genet Dev.* 2015;**35**:119–125. <https://doi.org/10.1016/j.gde.2015.11.003>
- Song Q, Zhang T, Stelly DM, Chen ZJ. Epigenomic and functional analyses reveal roles of epialleles in the loss of photoperiod sensitivity during domestication of allotetraploid cottons. *Genome Biol.* 2017;**18**(1):99. <https://doi.org/10.1186/s13059-017-1229-8>
- Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 2006;**34**(Web Server):W609–W612. <https://doi.org/10.1093/nar/gkl315>
- Takuno S, Gaut BS. Body-methylated genes in *Arabidopsis thaliana* are functionally important and evolve slowly. *Mol Biol Evol.* 2012;**29**(1):219–227. <https://doi.org/10.1093/molbev/msr188>
- Takuno S, Gaut BS. Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. *Proc Natl Acad Sci U S A.* 2013;**110**(5):1797–1802. <https://doi.org/10.1073/pnas.1215380110>
- Takuno S, Ran J-H, Gaut BS. Evolutionary patterns of genic DNA methylation vary across land plants. *Nat Plants.* 2016;**2**(2):15222. <https://doi.org/10.1038/nplants.2015.222>
- Tang H, Krishnakumar V, Bidwell S, Rosen B, Chan A, Zhou S, Gentzittel L, Childs KL, Yandell M, Gundlach H, et al. An improved genome release (version Mt4.0) for the model legume *Medicago truncatula*. *BMC Genomics* 2014;**15**(1):312. <https://doi.org/10.1186/1471-2164-15-312>
- The International Brachypodium Initiative. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* 2010;**463**(7282):763–768. <https://doi.org/10.1038/nature08747>
- Tran RK, Henikoff JG, Zilberman D, Ditt RF, Jacobsen SE, Henikoff S. DNA methylation profiling identifies CG methylation clusters in *Arabidopsis genes*. *Curr Biol.* 2005;**15**(2):154–159. <https://doi.org/10.1016/j.cub.2005.01.008>
- Turco GM, Kajala K, Kunde-Ramamoorthy G, Ngan C-Y, Olson A, Deshpande S, Tolkunov D, Waring B, Stelpflug S, Klein P, et al. DNA methylation and gene expression regulation associated with vascularization in *Sorghum bicolor*. *New Phytol.* 2017;**214**(3):1213–1229. <https://doi.org/10.1111/nph.14448>
- Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, et al. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 2006;**313**(5793):1596–1604. <https://doi.org/10.1126/science.1128691>
- Valliyodan B, Cannon SB, Bayer PE, Shu S, Brown AV, Ren L, Jenkins J, Chung CY-L, Chan T-F, Daum CG, et al. Construction and comparison of three reference-quality genome assemblies for soybean. *Plant J.* 2019;**100**(5):1066–1082. <https://doi.org/10.1111/tbj.14500>
- VanBuren R, Bryant D, Edger PP, Tang H, Burgess D, Challabathula D, Spittle K, Hall R, Gu J, Lyons E, et al. Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. *Nature* 2015;**527**(7579):508–511. <https://doi.org/10.1038/nature15714>
- VanBuren R, Wai CM, Colle M, Wang J, Sullivan S, Bushakra JM, Liachko I, Vining KJ, Dossett M, Finn CE, et al. A near complete, chromosome-scale assembly of the black raspberry (*Rubus occidentalis*) genome. *Gigascience* 2018;**7**(8):gy094. <https://doi.org/10.1093/gigascience/gy094>
- Van de Peer Y, Mizrahi E, Marchal K. The evolutionary significance of polyploidy. *Nat Rev Genet.* 2017;**18**(7):411–424. <https://doi.org/10.1038/nrg.2017.26>
- Verde I, Jenkins J, Dondini L, Micali S, Pagliarani G, Vendramin E, Paris R, Aramini V, Gazza L, Rossini L, et al. The Peach v2.0 release: high-resolution linkage mapping and deep resequencing improve chromosome-scale assembly and contiguity. *BMC Genomics* 2017;**18**(1):225. <https://doi.org/10.1186/s12864-017-3606-9>
- Wang H, Beyene G, Zhai J, Feng S, Fahlgren N, Taylor NJ, Bart R, Carrington JC, Jacobsen SE, Ausin I. CG gene body DNA methylation changes and evolution of duplicated genes in cassava. *Proc Natl Acad Sci U S A.* 2015;**112**(44):13729–13734. <https://doi.org/10.1073/pnas.1519067112>
- Wang W, Haberer G, Gundlach H, Gläßer C, Nussbaumer T, Luo MC, Lomsadze A, Borodovsky M, Kerstetter RA, Shanklin J, et al. The *Spirodela polyrhiza* genome reveals insights into its neotenus reduction fast growth and aquatic lifestyle. *Nat Commun.* 2014;**5**(1):3311. <https://doi.org/10.1038/ncomms4311>
- Wang J, Hossain MS, Lyu Z, Schmutz J, Stacey G, Xu D, Joshi T. SoyCSN: soybean context-specific network analysis and prediction based on tissue-specific transcriptome data. *Plant Direct.* 2019;**3**(9):e00167. <https://doi.org/10.1002/pld3.167>
- Wang Y, Li J, Paterson AH. MCScanX-transposed: detecting transposed gene duplications based on multiple colinearity scans. *Bioinformatics* 2013;**29**(11):1458–1460. <https://doi.org/10.1093/bioinformatics/btt150>
- Wang J, Marowsky NC, Fan C. Divergence of gene body DNA methylation and evolution of plant duplicate genes. *PLoS One* 2014;**9**(10):e110357. <https://doi.org/10.1371/journal.pone.0110357>
- Wang Y, Wang X, Lee T-H, Mansoor S, Paterson AH. Gene body methylation shows distinct patterns associated with different gene origins and duplication modes and has a heterogeneous relationship

- with gene expression in *Oryza sativa* (rice). *New Phytologist* 2013;**198**(1):274–283. <https://doi.org/10.1111/nph.12137>
- Wang L, Xie J, Hu J, Lan B, You C, Li F, Wang Z, Wang H.** Comparative epigenomics reveals evolution of duplicated genes in potato and tomato. *Plant J.* 2018;**93**(3):460–471. <https://doi.org/10.1111/tpj.13790>
- Wang X, Zhang Z, Fu T, Hu L, Xu C, Gong L, Wendel JF, Liu B.** Gene-body CG methylation and divergent expression of duplicate genes in rice. *Sci Rep.* 2017;**7**(1):1–11. <https://doi.org/10.1038/s41598-017-02860-4>
- Wang D, Zhang Y, Zhang Z, Zhu J, Yu J.** Kaks_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinformatics* 2010;**8**(1):77–80. [https://doi.org/10.1016/S1672-0229\(10\)60008-3](https://doi.org/10.1016/S1672-0229(10)60008-3)
- Wei T, Viliam S** R package “corrplot”: visualization of a correlation matrix (Version 0.84). R package “corrplot”: visualization of a correlation matrix (Version 0.84). Available from <https://github.com/taiyun/corrplot>.
- Wu GA, Terol J, Ibanez V, López-García A, Pérez-Román E, Borredá C, Domingo C, Tadeo FR, Carbonell-Caballero J, Alonso R, et al.** Genomics of the origin and evolution of Citrus. *Nature* 2018;**554**(7692):311–316. <https://doi.org/10.1038/nature25447>
- Xu S, Brockmüller T, Navarro-Quezada A, Kuhl H, Gase K, Ling Z, Zhou W, Kreitzer C, Stanke M, Tang H, et al.** Wild tobacco genomes reveal the evolution of nicotine biosynthesis. *Proc Natl Acad Sci U S A.* 2017;**114**(23):6133–6138. <https://doi.org/10.1073/pnas.1700073114>
- Xu C, Nadon BD, Kim KD, Jackson SA.** Genetic and epigenetic divergence of duplicate genes in two legume species. *Plant Cell Environ.* 2018;**41**(9):2033–2044. <https://doi.org/10.1111/pce.13127>
- Potato Genome Sequencing Consortium, Xu X, Pan S, Cheng S, Zhang B, Mu D, Ni P, Zhang G, Yang S, Li R, et al.** Genome sequence and analysis of the tuber crop potato. *Nature* 2011;**475**(7355):189–195. <https://doi.org/10.1038/nature10158>
- Xue H, Wang S, Yao J-L, Deng CH, Wang L, Su Y, Zhang H, Zhou H, Sun M, Li X, et al.** Chromosome level high-density integrated genetic maps improve the *Pyrus bretschneideri* “DangshanSuli” v1.0 genome. *BMC Genom.* 2018;**19**(1):1–13. <https://doi.org/10.1186/s12864-018-5224-6>
- Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E, et al.** Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* 2005;**21**(5):650–659. <https://doi.org/10.1093/bioinformatics/bti042>
- Yang Z, Bielawski JP.** Statistical methods for detecting molecular adaptation. *Trends Ecol Evol (Amst).* 2000;**15**(12):496–503. [https://doi.org/10.1016/S0169-5347\(00\)01994-7](https://doi.org/10.1016/S0169-5347(00)01994-7)
- Yang R, Jarvis DE, Chen H, Beilstein MA, Grimwood J, Jenkins J, Shu S, Prochnik S, Xin M, Ma C, et al.** The reference genome of the halophytic plant *Eutrema salsugineum*. *Front Plant Sci.* 2013;**4**:46. <https://doi.org/10.3389/fpls.2013.00046>
- Yang Y, Tang K, Datsenka TU, Liu W, Lv S, Lang Z, Wang X, Gao J, Wang W, Nie W, et al.** Critical function of DNA methyltransferase 1 in tomato development and regulation of the DNA methylome and transcriptome. *J Integr Plant Biol.* 2019;**61**(12):1224–1242. <https://doi.org/10.1111/jipb.12778>
- Zemach A, McDaniel IE, Silva P, Zilberman D.** Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* 2010;**328**(5980):916–919. <https://doi.org/10.1126/science.1186366>
- Zhang J.** Evolution by gene duplication: an update. *Trends Ecol Evol (Amst).* 2003;**18**(6):292–298. [https://doi.org/10.1016/S0169-5347\(03\)00033-8](https://doi.org/10.1016/S0169-5347(03)00033-8)
- Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW-L, Chen H, Henderson IR, Shinn P, Pellegrini M, Jacobsen SE, et al.** Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. *Cell* 2006;**126**(6):1189–1201. <https://doi.org/10.1016/j.cell.2006.08.003>
- Zhao XP, Si Y, Hanson RE, Crane CF, Price HJ, Stelly DM, Wendel JF, Paterson AH.** Dispersed repetitive DNA has spread to new genomes since polyploid formation in cotton. *Genome Res.* 1998;**8**(5):479–492. <https://doi.org/10.1101/gr.8.5.479>