Research Article

Preferential retention of the slowly evolving gene in pairs of duplicates in angiosperm genomes

Lin-Bo Wang1,2, Hong Ma3, and Juan Lin1*

1State Key Laboratory of Genetic Engineering, School of Life Sciences, Institute of Plant Biology, Fudan University, Shanghai 200433, China
2Institute of Science and Technology for Brain-inspired Intelligence, Fudan University, Shanghai 200433, China
3Department of Biology, The Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA 16802, USA

*Author for correspondence. E-mail: linjuan@fudan.edu.cn

Received 5 August 2020; Accepted 30 November 2020; Article first published online 3 December 2020

Abstract Gene duplication provides raw material for functional innovation, but gene duplicability varies considerably. Previous studies have found widespread asymmetrical sequence evolution between paralogs. However, it remains unknown whether the rate of evolution among paralogs affects their propensity of being retained after another round of whole-genome duplication (WGD). In this study, we investigated gene groups that have experienced two successive WGDs to determine which of two older duplicates with different evolutionary rates was more likely to retain both younger duplicates. To uncouple the measurement of evolutionary rates from any assignment of duplicate or singleton status, we measured the evolutionary rates of singleton genes in out-lineages but classified these singleton genes according to whether they are retained or not in a crown group of species. We found that genes that retained younger duplicates in the crown group of genomes were more constrained prior to the younger duplication event than those that failed to leave duplicates. In addition, we also found that the retained clades have more genes in out-lineages. Subsequent analyses showed that genes in the retained clades were expressed more broadly and highly than genes in the singleton clades. We concluded that the set of repeatedly retained genes after two WGDs is biased toward slowly evolving genes in angiosperms, suggesting that the potential of genes for both functional conservation and divergence likely affects their propensity of being retained after WGD in angiosperms.

Key words: angiosperm, evolutionary rate, gene duplication, preferential retention, subfunctionalization.

1 Introduction

Gene duplication is the primary source of new genes during evolution and is consequently a vital source of genetic novelty. Whole-genome duplications (WGDs) generate a copy of all genes and are associated with organismal diversification and the acquisition of novel features (Rensing, 2014). Angiosperms have experienced many rounds of WGDs, in the most recent common ancestor (MRCA) of angiosperms and at the levels of the order or family (Fig. 1A) (Jiao et al., 2011; Vanneste et al., 2014; Huang et al., 2016b; Xiang et al., 2017; Ren et al., 2018). Recent WGD events also occurred independently within many plant groups, such as Brassicaceae (Barker et al., 2009), Fabaceae (Cannon et al., 2015), Poales (McKain et al., 2016), and Poaceae (Ren et al., 2018). In addition, more recent polyploidizations have been detected in Brassica rapa var. pekinensis (L.) Kita, (Wang et al., 2011), Glycine max (L.) Merr (Schmutz et al., 2010), and Zea mays L. (Ren et al., 2018). However, most duplicates obtained from WGDs were silenced and then lost within a few million years (Li et al., 2016; Ren et al., 2018).

Those that are retained may meet one of several fates: (i) functional conservation, which is beneficial due to the increased amount of ancestral gene product (Ohno, 1970); (ii) subfunctionalization, which divides the ancestral function among the two copies (Lynch & Conery, 2000); (iii) neofunctionalization, which entails the acquisition of a new function by one copy while the other retains the original function (Assis & Bachtrog, 2013; Turetzek et al., 2016); or (iv) specialization, a combination of (ii) and (iii) (Jiang & Assis, 2017). Both (iii) and (iv) suggest that two duplicates could evolve at different rates.

Indeed, widespread asymmetrical evolution has been supported by substantial empirical evidence (Conant & Wagner, 2003; Pich & Kondrashov, 2014; Zhou et al., 2014; Wang et al., 2017). Approximately 20%–30% of the duplicates in Saccharomyces cerevisiae (Desm.) Meyen, Schizosaccharomyces pombe Lindn, Drosophila melanogaster Meigen, and Caenorhabditis elegans Maupas show traces of having evolved asymmetrically (Conant & Wagner, 2003). Similarly, approximately one-fifth of pairs of recently duplicated human genes have evolved asymmetrically (Panchin
et al., 2010). In a previous study that focused on recent duplicates in rodents, it was found that acceleration in the rate of evolution after gene duplication is widespread across genes but is generally limited to only one copy (Pegueroles et al., 2013). In land plants, most duplicates have asymmetrical rates of evolution or significant differences in the substitution rates (Carretero-Paulet & Fares, 2012). Asymmetrical expression evolution, where one WGD paralog has low or no expression in all tissues, is also prevalent between duplicates (Assis & Bachtrog, 2013). Altogether, many studies have reported that asymmetrical evolution is common in gene duplication.

Another question regarding gene duplication is whether genes with different rates of sequence evolution show differential propensities for the retention of subsequent duplicates (i.e., gene duplicability). Although several studies have investigated the possible relationships between rates of coding sequence evolution and gene duplicability (Davis & Petrov, 2004; Brunet et al., 2006; He & Zhang, 2006; Liang & Li, 2007; Semon & Wolfe, 2008; O’Toole et al., 2018), they have reached different conclusions. For example, in yeast genomes, less-important genes have higher rates of successful duplication (He & Zhang, 2006), suggesting that more rapidly evolving genes have a higher chance of retaining their duplicates. In primate genomes, genes with family-size conservation evolve more slowly than those without (Chen et al., 2010), and rapidly evolving genes tend to be without family-size conservation (O’Toole et al., 2018). However, others have argued that genes that have more important biological functions and whose coding sequences are under stronger functional constraints are retained more often (Liang & Li, 2007). Studies of eukaryotic genomes found that duplicates encoding more conserved proteins are retained more often (Davis & Petrov, 2004), and duplicates also evolve more slowly than genes in singleton clades (Jordan et al., 2004). However, in plants, the relationship between the rates of coding sequence evolution and gene duplicability remain elusive.

Angiosperms have experienced many rounds of WGDs (Fig. 1A), and these large-scale duplications have generated numerous pairs of duplicates, which provide an excellent opportunity to analyze the differential sequence evolution of

Fig. 1. Species phylogeny and the approach used to estimate the rate of evolution for retained and singleton clades. A, The species phylogeny in this study was inferred from the phylogeny of conserved nuclear genes (Zeng et al., 2014, 2017; Huang et al., 2016a). Genome duplication and triplication events are marked according to previous reports (Jiao et al., 2011; Vanneste et al., 2014; Huang et al., 2016b; Ren et al., 2018) and are depicted by stars. Species names given in red are used as family-level out-lineage species. B, Genes that experienced two successive duplications and the orthologs of these genes in the out-lineage. We were interested in identifying genes that are singletons in species S1, S2, and S3 and in sorting those genes into Group A (retained clade) and Group B (singleton clade), according to their status in the S4 genome. For all genes in Group A (light green) and Group B (blue) in S4, their unduplicated orthologs (relative to younger duplication) in species S1, S2, and S3 were identified. C, Calculating the proxy for the ancestral evolutionary rate. The rate of evolution in the orthologous pairs in out-lineage species (S1, S2, and S3) was used as a proxy for the ancestral rate of evolution in the S4 lineage, independent of the younger duplication event in molecular evolution.
Wang et al.

2 Material and Methods

2.1 Data collection

Sequence data collection of 30 species and phylogenetic analyses were carried out exactly as described in a previous article (Wang et al., 2019). The sequence of *Vitis vinifera* was downloaded from the public database Phytozome version 12.0 (https://phytozome.jgi.doe.gov/pz/portal.html) (Goodstein et al., 2012). The transcript and protein expression data for *Arabidopsis thaliana* (L.) Heynh. were obtained from the multi-omics atlas given in a previously published paper (Mergner et al., 2020). mRNA quantities are displayed as transcripts per million. Protein intensities are intensity-based absolute quantification values (Mergner et al., 2020). The level of expression of each gene was estimated using the median value of all 30 tissues. The tissue specificity was measured with the index $\tau$ (Yanai et al., 2005):

$$\tau = 1 - \frac{\sum_{i=1}^{N}(1 - x_i)}{N - 1}$$

where $N$ is the number of tissues, and $x_i$ is a normalized expression level for each tissue, divided by the highest level of expression across the $N$ tissues. The index $\tau$ ranges from 0 to 1, and higher values indicate higher specificity. If a gene was expressed in only one tissue, it was assigned $\tau = 1$, and if it was expressed equally in all tissues, it was assigned $\tau = 0$.

2.2 Gene tree construction and reconciliation

We collected the protein-coding sequences for 30 sequenced angiosperm genomes that can stand in for major angiosperm lineages (Fig. 1A), and we classified them into orthogroups using OrthoFinder (Emms & Kelly, 2015). Of the 177 988 clusters that were obtained using OrthoFinder, 12 046 contained genes from at least five different species, which were included in subsequent analyses. Of these, 74 clusters with more than 500 genes were removed because of the enormous computational resources required by the large number of sequences.

We implemented a gene tree-species tree reconciliation pipeline to automatically construct phylogenetic gene trees for all clusters to obtain estimates of duplication and speciation events along the gene tree and to test whether these trees could be traced back to a single ancestral angiosperm gene. The pipeline incorporated a procedure for gene tree improvement using TreeFix (Wu et al., 2013), a gene tree-species tree reconciliation algorithm that carries out topological changes to gene trees and searches for alternative topologies that minimize the duplication/loss cost, with a likelihood that is not statistically worse than that of the maximum likelihood topology. The tree reconciliation algorithm used a predefined species tree (Fig. 1A) drawn from recent studies of angiosperm phylogeny to infer all of the necessary gene gains and losses to explain a given gene tree topology (Zeng et al., 2014, 2017; Huang et al., 2016a). The phylogenetic analyses were undertaken as presented previously (Wang et al., 2019). Briefly, sequence alignments based on multiple protein sequences were produced using Multiple Alignment using Fast Fourier Transform (MAFFT) version 7.273 (Katoh & Standley, 2013) and were trimmed with trimAl (Capella-Gutierrez et al., 2009), using the options “–gt 0.1 -resoverlap 0.75 -seqoverlap 80.” The processed multiple sequence alignments were fed into the IQ-TREE (Nguyen et al., 2015) to construct phylogenetic gene trees, using the amino acid substitution (LG+F+G) model (Le & Gascuel, 2008). Then codon-based coding sequence alignments generated from the sequence alignments of amino acids were back translated using TreeBeST version 1.9.2 (Villella et al., 2009) and used for the analyses with TreeFix (Wu et al., 2013) with the default reconciliation model (duplication/loss cost), but increasing the number of iterations to 1000. If a reconstructed gene tree generated by TreeFix indicated a duplication event at a node before the MRCA of angiosperms, we split the phylogenetic tree into subtrees using ETE3 (Huerta-Cepas et al., 2016), a Python framework for phylogenetic analyses.
2.3 Identification of duplicate and singleton genes and their orthologs

Duplication and speciation nodes were detected by ETE3 (Huerta-Cepas et al., 2016), using a reconstructed gene tree generated by TreeFix, in which duplication and separation nodes strictly followed the species tree topology (Zeng et al., 2014, 2017; Huang et al., 2016a). In this study, only duplication nodes that exceeded a duplication consistency score of 0.2 were considered in subsequent analyses, following the procedure of a previous study (Villella et al., 2009). The consistency score is the intersection of the number of species post-duplication divided by the union (Villella et al., 2009), which is a measure of the plausibility of the duplication events inferred. By detecting evolutionary events, orthologous and paralogous relationships among genes can also be inferred.

2.4 Measurement of proxy-ancestral evolutionary rate

For each of the orthologous pairs, the protein sequences were aligned using MAFFT version 7.273 (Katoh & Standley, 2013) and then converted into nucleotide alignments using TreeBeST version 1.9.2 (Villella et al., 2009). $K_o$, $K_s$, and $\omega (K_s/K_o)$ were estimated using CodeML software from the PAML package (version 4.8) (Yang, 2007), based on codon sequence alignments using the GY model, with stationary codon frequencies empirically estimated by the F3x4 model. Using the alignment of out-lineage orthologous pairs of genes, we prevented gene duplication from affecting inferences of evolutionary rates. Orthologous pairs with $K_o > 5$ or $K_s < 0.01$ were excluded from further analyses, which could denote either misalignment or potential sequence saturation.

3 Results

3.1 Evolutionary rates for orthologous singleton genes are highly correlated in related lineages

In all, 30 species with genome sequences (22 eudicots, seven monocots, and one basal angiosperm) (Fig. 1A) were selected to represent the major angiosperm lineages. Duplicated and singleton genes were inferred from gene trees (see Section 2 for details). Then we undertook a systematic search for genes that were singleton at the base of the eudicots tree in Aquilegia coerulea L., Carica papaya L., Theobroma cacao L., Fragaria vesca L., and Vitis vinifera L., which lacked evidence of WGDs after divergence from other eudicots (hereafter, family-level out-lineage species (FLOS)). Family-level out-lineage species are not similar to Brassicaceae and Fabaceae species, whose own MRCA experienced at least one ancestral WGD. We identified genes for which we could be confident were singletons at the base of the eudicot tree. To assess whether the rates of evolution for orthologous singleton genes were correlated in angiosperms, $K_o$ values were calculated for these orthologous pairs in a pairwise mode. Overall, the $K_o$ values were highly positively correlated across orthologous pairs (Spearman's rank correlation, $R > 0.5$, $P < 0.001$) (Fig. S1), suggesting that they generally had similar rates of evolution. We also repeated the entire analyses in Brassicaceae, Fabaceae, and Poaceae-level orthologous gene pairs, and the results remained unchanged and strongly statistically significant (Spearman's rank correlation, $R > 0.5$, $P < 1.0e-16$) (Figs. S2–S4). From these results, we reasoned that the $K_o$ values between the members of each representative pair would be a good proxy for the rate of evolution of the duplicates in a way that would be unaffected by the process of duplication.

3.2 Genes in retained clades evolve at slower rates than those in singleton clades

From these species (Fig. 1A), we identified a large number of gene duplications in the core eudicots, Brassicaceae, Fabaceae, Poaceae, Brassica rapa, Glycine max, and Zea mays. Because we focused on genes that experienced two successive duplications (old and young WGD) (Fig. 1B), we chose the following five gene datasets: core eudicots and Brassicaceae duplicates (dataset 1) (Fig. 2A), core eudicots and Fabaceae duplicates (dataset 2) (Fig. 2C), Brassicaceae and B. rapa duplicates (dataset 3) (Fig. 3A), Fabaceae and G. max duplicates (dataset 4) (Fig. 3C), and Poaceae and Z. mays duplicates (dataset 5) (Fig. 3E). Then we classified the genes of these five datasets into either Group A or Group B based on their duplication status after young WGD. Three copies generated by the $\gamma$ triplication (core eudicot-wide WGD) event were also classified into two groups, with one group (Group A or Group B) containing two old clades. For three or four copies generated by the $\alpha\beta$ duplication event, only copies generated by the second-round old WGD were included in this study, which requires that both Group A and Group B contained one old clade. For each of two successive duplications, Group A always contained more genes than Group B in crown lineages or species after the younger WGD (Fig. 1B). For clarity, Group A is called the retained clade and Group B is the singleton clade.

First, we analyzed dataset 1 genes (core eudicots and Brassicaceae duplicates) (Fig. 2A). Genes from dataset 1 were partitioned into the retained clade and the singleton clade, based on their status in Brassicaceae. Because we were interested in investigating the evolutionary rate of the retained and singleton clades, independent of the potential effects of the Brassicaceae-level duplication event, we measured the amount of evolution ($K_o$ and $\omega$) in two FLOS, T. cacao and C. papaya, and we used the results as a proxy for the ancestral evolutionary rate of the Brassicaceae genes. We observed significant differences between the retained clade and the singleton clade in $K_o$ and the $\omega$ values (Mann–Whitney U-test, $P < 0.05$) (Fig. 2B). The retained clade showed lower $K_o$ and $\omega$ values, although these measurements were based on the comparison between orthologs from the T. cacao and C. papaya sequences, where the genes had not experienced Brassicaceae-level WGD. We also carried out another set of comparisons, using both V. vinifera and T. cacao, and A. coerulea and T. cacao (Fig. 5S). The retained clades have significantly lower $\omega$ values (between orthologs from V. vinifera and T. cacao (Fig. 5S$A$)) and $K_o$ values (between orthologs from V. vinifera and T. cacao (Fig. 5S$B$)) than those between similar orthologs in singleton clades (Mann–Whitney U-test, $P < 0.05$). The decrease in statistical power in these comparisons could be due to the relatively long divergence times of the two pairs of species and limited gene number. We also repeated the whole set of analyses for the genes in dataset 2 (Fig. 2C) and measured the amount of evolution ($K_o$ and $\omega$) between orthologs from F. vesca and T. cacao. The results remained statistically significant.
Fig. 2. Boxplots showing \( K_a, K_s \), and \( \omega (K_a/K_s) \) values for the retained clade (Group A) and the singleton clade (Group B) gene in the eudicot gene set. All genes had singleton status relative to family-level whole-genome duplications, and the rates of evolution were calculated by comparison with orthologous pairs in out-lineage species. The values were compared using the Mann–Whitney U-test (unpaired), and \( P \)-values are shown above each pair of boxplots. **A, B**, Genes in out-lineage species (i.e., *Theobroma cacao* (Thca) and *Carica papaya* (Capa)) are classified into retained and singleton clades, according to their status in the Brassicaceae genomes (\( n = 382 \)). **C, D**, Genes in out-lineage species (i.e., *Fragaria vesca* (Frve) and *T. cacao*) are classified into singleton and retained clades, according to their status in the Fabaceae genome (\( n = 245 \)). NS, not significant.

(Mann–Whitney U-test, \( P < 0.05 \)) (Fig. 2D). In addition, significantly lower \( K_a \) values between orthologs from *A. coerulea* and *F. vesca* of the retained clades were also observed in the comparisons to singleton clades (Fig. S6).

As the \( \gamma \) WGDs and the family-level WGD (i.e., the Brassicaceae WGD (\( \alpha /\beta \)), the Poaceae WGD (\( \rho \)), and the Fabaceae WGD) are relatively old, to reveal the patterns of relatively younger post-duplication retention biases in datasets 3–5, we compared the evolutionary rate of the retained clade and the singleton clade, partitioned by species-level WGD events (i.e., the *B. rapa* WGD, the *G. max* WGD, and the *Z. mays* WGD) (Fig. 3). We also found that the evolutionary rate of retained clades was statistically significantly lower than the singleton clades for all pairs of comparisons mentioned above (Mann–Whitney U-test, \( P < 0.05 \)) (Figs. 3, S7–S9). A paired-sample Mann–Whitney U-test was used in the comparisons between retained clades and singleton clades in datasets 3–5. Together with results described above, we concluded that the retained clade experienced a generalized higher constraint than the singleton clade, observed as slower sequence evolution. This is consistent with the results of a previous study that found preferential duplication of conserved proteins in eukaryotic genomes (Davis & Petrov, 2004).

### 3.3 Genes in retained clades are expressed more widely and highly than those in singleton clades

The level of gene expression is related to the rate of sequence evolution (Alvarez-Ponce & Fares, 2012; Zhang & Yang, 2015). To test whether genes in retained clades were expressed more broadly or highly than genes in singleton clades, we compared the gene expression pattern of these two categories of *Arabidopsis thaliana* genes. These genes were partitioned into a retained clade and singleton clade, based on their status in *B. rapa*. We found that *A. thaliana* genes in the retained clade were expressed more broadly (lower tissue specificity) and highly (median value) than genes in the singleton clade at the transcript and protein level (Mann–Whitney U-test, \( P < 0.05 \)) (Fig. 4), supporting the model of the preferential retention of genes with broader or higher expression levels.

### 3.4 Retained clades contain more genes than singleton clades in out-lineage species

If a retained clade has greater duplicability than a singleton clade, it can also be expected to be more likely to be retained in related out-lineage species. To test this hypothesis, we compared the number of genes in retained clades and singleton clades in FLOS. First, we
partitioned paralogous clades generated by $\gamma$ into two classes, namely, the retained Brassicaceae clade and the singleton Brassicaceae clade. We measured the number of genes in *Mimulus guttatus* DC, *Lactuca sativa* L., *Cucumis sativus* L., and *Populus trichocarpa* (Torr. & A.Gray), all of which showed evidence of WGD events after divergence from other angiosperms (Ren et al., 2018). The results showed that retained clades consistently had more genes than singleton clades (Mann–Whitney U-test, $P < 0.05$) (Fig. 5A). We repeated this entire set of analyses for datasets 2–5, and the retained clades consistently had more genes than singleton clades (Mann–Whitney U-test, $P < 0.05$) (Figs. 5B, S10–S13), suggesting that retained clades were more likely to be retained after WGD. Because some of these out-lineage species lack evidence of WGD events after their divergence from other angiosperms, these results also suggest that genes with lower rates of evolution are inherently more likely to be retained after small-scale duplication, or genes with higher rates of evolution are inherently more likely to be lost.

4 Discussion

Previous studies have shown that some genes, the so-called duplicable genes, are more likely to be retained after WGDs (Li et al., 2016; Qiao et al., 2019; Wang et al., 2019). In this...
study, we asked a different question: among WGD-derived paralogs, which gene was more likely to be retained after a second round of WGD? The approach that was used to identify duplicates in one organism or family and to obtain evolutionary rate measurements from two out-lineage species (Fig. 1) allowed the comparison of the evolutionary rate of genes between those retained to that those that were not. Importantly, this allowed us to make this comparisons while avoiding the confounding effects that the duplication itself has on the rates of sequence evolution (Davis & Petrov, 2004; He & Zhang, 2006; Li et al., 2006; O’Toole et al., 2018). Our data showed that previous duplicates that have evolved slowly, presumably reflecting high functional constraints, were preferentially retained after second-round WGD across angiosperms. Because different functional classes of genes have different evolutionary rates in a genome (Koonin & Wolf, 2010), we compared the evolutionary rate of retained clades and singleton clades in a pairwise mode free from functional classes bias. Due to the long periods of evolutionary time we examined (i.e., $\gamma$ shared by core eudicots, various family-level WGD events, and various species-level WGD events), the retention biases were more important in this study than in others that have examined shorter lengths of time (O’Toole et al., 2018). In addition, ascertaining the magnitude of any effect would require forward simulations.

These results provide a consistent framework for considering several previous findings, as follows: highly connected proteins tend to have higher gene duplicability in humans (Liang & Li, 2007); gene duplicates maintained after WGDs exhibit the strongest purifying selection in teleost fishes (Brunet et al., 2006); the bias of the set of duplicate genes has important implications for genome-scale studies (Davis & Petrov, 2004); gene duplicability in core genes is highly consistent across all angiosperms (Li et al., 2016); and slow-evolving genes after allopolyploidization in Xenopus laevis Daudin show preferential subfunctionalization (Semon & Wolfe, 2008). In addition, we previously found that duplicates of widely expressed and slowly evolving AP2/ERF genes were retained repeatedly and relatively often and continue to experience functional constraints (Wang et al., 2019). All of these results support a model where a subset of the genome is strongly constrained in sequence evolution but is easily retained after WGD.

One possible explanation for these results is that a slowly evolving (and functionally important) gene might be expressed at high level, such that a duplicated slowly evolving gene might have a better chance of survival than a duplicated rapidly evolving one (Li et al., 2016). Another possible explanation is that a slowly evolving gene, which usually has multiple functions, might have a greater potential of functional diversification (e.g., subfunctionalization) for the duplicates. Theoretical studies of population genetics also predict that subfunctionalization is a much more common mechanism of the preservation of duplicate genes than neofunctionalization (degenerative mutations are much more frequent than beneficial changes) (Force et al., 1999). These observations led us to propose that among WGD-derived paralogs, slowly evolving ones were more easily subfunctionalized (or more likely to experience specialization) and therefore were more likely to be retained long after the WGD, as illustrated in a model of the evolution of duplicable genes after WGD in angiosperms shown in Fig. 6.

In this model, two duplicates derived from one gene accumulate sequence differences, but more so in the rapidly evolving one than in the slowly evolving one. After second-round WGD, copies from the slowly evolving gene were more easily retained. One explanation is that it increased the amount of the slowly evolving gene product, which might need to be expressed at a high level (Li et al., 2016). An alternative explanation is that the slowly evolving gene, which tends to have multiple functions (being expressed in numerous tissues or encoding multidomain proteins), were more easily subfunctionalized, resulting in functional diversification (Liang & Li, 2007; Alvarez-Ponce & Fares, 2012; Zhang & Yang, 2015). Then, newly subfunctionalized duplicates, which could acquire new functions, facilitate a more sophisticated regulation of signaling networks or adaptation to new environmental conditions. By contrast, the rapidly evolving gene, which is more likely already
Specialized and is not expressed at high level, tends to have low potential to be further subfunctionalized after duplication. For these reasons, the additional copy of the rapidly evolving gene after a second round WGD is less likely to be retained. Thus, among WGD-derived paralogs, the persistent conservation or subfunctionalization of the slowly evolving duplicate might be a mechanism for repeated gene retention after WGDs. The evolutionary scenario presented here might also be true for WGD-derived duplicates in other lineages.

**Fig. 5.** Boxplots showing the number of genes in retained clades and singleton clades in family-level out-lineage species. Numbers of genes greater than 10 are not shown. Values were compared using the Mann–Whitney U-test (unpaired), and P-values are shown above each pair of boxplots. A, Genes were separated into retained clades and singleton clades according to their status in the Brassicaceae genome (retained clade, n = 912; singleton clade, n = 1144). B, Genes were separated into retained and singleton clades according to their status in the Fabaceae genome (retained clade, n = 809; singleton clade, n = 980). Cusa, *Cucumis sativus*; Lasa, *Lactuca sativa*; Migu, *Mimulus guttatus*; Potr, *Populus trichocarpa*.

Duplicates that retained again evolve slowly.
References

Alvarez-Ponce D, Fares MA. 2012. Evolutionary rate and duplicability in the Arabidopsis thaliana protein–protein interaction network. Genome Biology and Evolution 4: 1263–1274.

Assis R, Bachtrog D. 2013. Neofunctionalization of young duplicate genes in Drosophila. Proceedings of the National Academy of Sciences USA 110: 17409–17414.

Barker MS, Vogel H, Schranz ME. 2009. Paleopolyploidy in the Brassicaceae: Analyses of the cleome transcriptome elucidate the history of genome duplications in Arabidopsis and other Brassicaceae. Genome Biology and Evolution 1: 391–399.

Bromham L, Penny D. 2003. The modern molecular clock. Nature Reviews Genetics 4: 216–224.

Brunet FG, Roest Crollius H, Paris M, Aury JM, Gibert P, Jaillon O, Laudet V, Robinson-Rechavi M. 2006. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. Molecular Biology and Evolution 23: 1808–1816.

Cannon SB, McKain MR, Harkest A, Nelson MN, Dash S, Deyholos MK, Peng Y, Joyce B, Stewart CN, Jr, Rolf M, Kutchan T, Tan X, Chen C, Zhang Y, Carpenter E, Wong GK, Doyle JJ, Leebens-Mack J. 2015. Multiple polyploidy events in the early radiation of Brassicales: Analyses of the cleome transcriptome elucidate the history of genome duplications in Arabidopsis and other Brassicaceae. Genome Biology and Evolution 1: 391–399.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. Trimal: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25: 1927–1937.

Carretero-Paulet L, Fares MA. 2012. Evolutionary dynamics and functional specialization of plant paralogs formed by whole and small-scale genome duplications. Molecular Biology and Evolution 29: 3541–3551.

Chen C, Zhang Y, Carpenter E, Wong GK, Doyle JJ, Leebens-Mack JH. 2016b. Gene duplicability of core genes is highly consistent across all angiosperms. The Plant Cell 28: 326–344.

Chiang X, Assis R. 2017. Natural selection drives rapid functional evolution of young Drosophila duplicate genes. Molecular Biology and Evolution 34: 3089–3098.

Jiang Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, Soltis DE, Clifton SW, Scharbaum SE, Schuster SC, Ma H, Leebens-Mack J, dePamphilis CW. 2011. Ancestral polyploidy in seed plants and angiosperms. Nature 473: 97–100.

Jordan IK, Wolf YI, Koonin EV. 2004. Duplicated genes evolve slower than singletons despite the initial rate increase. BMC Evolutionary Biology 4: 22.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Molecular Biology and Evolution 30: 772–780.

Koonin EV, Wolf YI. 2010. Constraints and plasticity in genome and molecular-phenome evolution. Nature Reviews Genetics 11: 487–498.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. Molecular Biology and Evolution 25: 1307–1320.

Li L, Huang Y, Xia X, Sun Z. 2006. Preferential duplication in the sparsely part of yeast protein interaction network. Molecular Biology and Evolution 23: 2467–2473.

Li Z, Defoort J, Tasdighian S, Maere S, Van de Peer Y, De Smet R. 2016. Gene duplicability of core genes is highly consistent across all angiosperms. The Plant Cell 28: 326–344.

Liang H, Li WH. 2007. Gene essentiality, gene duplicability and protein connectivity in human and mouse. Trends in Genetics 23: 375–378.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science 290: 1151–1155.

McKain MR, Tang H, McNeal JR, Ayyampalayam S, Davis JJ, dePamphilis CW, Givnish TJ, Pires JC, Stevenson DW, Leebens-Mack JH. 2016. A phylogenomic assessment of ancient polyploidy and genome evolution across the Poales. Genome Biology and Evolution 8: 1150–1164.

Megrner J, Frejno M, List M, Papacek M, Chen X, Chaudhary A, Samaras P, Richter S, Shikata H, Messerer M, Lang D, Altmann S, Cyprys P, Zolg DP, Mathiesen T, Bantscheff M, Hazarika RR, Schmidt T, David C, Dunkel A, Hofmann T, Sprunk S, Kuster B. 2020. Mass-spectrometry-based draft of the Arabidopsis proteome. Nature 579: 409–414.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-tree: A rapidly and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and Evolution 32: 268–274.

O’Toole AN, Hurst LD, McLean SG, McFadden J, Stewart CN, Jr, Rolf M, Ruzicka DR, Wafafa E, Wickett NJ, Wu X, Zhang Y, Wang J, Zhang Y, Carpenter EJ, Deyholos MK, Kutchan TM, Chanderbali AS, Soltis PS, Stevenson DW, McCombie R, Pires JC, Wong GK, Soltis DE, dePamphilis CW. 2012. A genome triplication associated with early diversification of the core eudicots. Genome Biology 13: R3.
Pegueroles C, Laurie S, Alba MM. 2013. Accelerated evolution after gene duplication: A time-dependent process affecting just one copy. *Molecular Biology and Evolution* 30: 1830–1842.

Pich IRO, Kondrashov FA. 2014. Long-term asymmetrical acceleration of protein evolution after gene duplication. *Genome Biology and Evolution* 6: 1949–1955.

Qiao X, Li Q, Yin H, Qi K, Li L, Wang R, Zhang S, Paterson AH. 2019. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome Biology* 20: 38.

Ren R, Wang H, Guo C, Zhang N, Zeng L, Chen Y, Ma H, Qi J. 2018. Widespread whole genome duplications contribute to genome complexity and species diversity in angiosperms. *Molecular Plant* 11: 414–428.

Rensing SA. 2014. Gene duplication as a driver of plant morphogenetic evolution. *Current Opinion in Plant Biology* 17: 43–48.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, Gill N, Joshi T, Libault M, Sethuraman A, Zhang XC, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC, Jackson SA. 2010. Genome sequence of the *Xenopus laevis* boundary. *Genome Research* 20: 38–49.

Semen M, Wolfe KH. 2008. Preferential subfunctionalization of slow-evolving genes after allopolyploidization in *Xenopus laevis*. *Proceedings of the National Academy of Sciences USA* 105: 8333–8338.

Turetzek N, Pechmann M, Schomburg C, Schneider J, Prpic NM. 2016. Neofunctionalization of a duplicate dachshund gene underlies the evolution of a novel leg segment in arachnids. *Molecular Biology and Evolution* 33: 109–121.

Vanneste K, Baele G, Maere S, Van, de Peer Y. 2014. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary. * Genome Research* 24: 1334–1347.

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. *EnsemblCompar* *eGeneTrees*: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research* 19: 327–335.

Wang L, Ma H, Lin J. 2019. Angiosperm-wide and family-level analyses of *ap2/erf* genes reveal differential retention and sequence divergence after whole-genome duplication. *Frontiers in Plant Science* 10: 196.

Wang M, Tu L, Lin M, Lin Z, Wang P, Yang Q, Ye Z, Shen C, Li J, Zhang L, Zhou X, Nie X, Li Z, Guo K, Ma Y, Huang C, Jin S, Zhu L, Yang X, Min L, Yuan D, Zhang Q, Lindsey K, Zhang X. 2017. Asymmetric subgenome selection and cis-regulatory divergence during domestication. *Nature Genetics* 49: 579–587.

Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun JH, Bancroft I, Cheng F, Huang S, Li X, Hua W, Wang J, Wang X, Freeling M, Pires JC, Paterson AH, Chalhoub B, Wang B, Hayward A, Sharpe AG, Park BS, Weißhaar B, Liu B, Li B, Liu B, Tong C, Song C, Duran C, Peng C, Geng C, Koh C, Lin C, Edwards D, Mu D, Shen D, Soumpourou E, Li F, Fraser F, Conant G, Lassalle G, King GJ, Bonnema C, Tang H, Wang H, Belcram H, Zhou H, Hirakawa H, Abe H, Guo H, Wang H, Jin H, Parkin IA, Batley J, Kim JS, Just L, Li J, Xu J, Deng J, Kim JA, Li J, Yu J, Meng J, Wang J, Min J, Pouliain J, Wang J, Hatakeyama K, Wu K, Wang L, Fang L, Trick M, Links MG, Zhao M, Jin M, Ramchiray N, Drou N, Berkman PJ, Cai Q, Huang Q, Li R, Tabata S, Cheng S, Zhang S, Zhang S, Huang S, Sato S, Sun S, Kwon SJ, Choi SR, Lee TH, Fan W, Zhao X, Tan X, Xu X, Wang Y, Qiu Y, Yin Y, Li Y, Du Y, Liao Y, Lim Y, Narusaka Y, Wang Y, Wang Z, Li Z, Wang Z, Xiong Z, Zhang Z, Brassica rapa Genome Sequencing Project Consortium. 2011. The genome of the mesopolyploid crop species *Brassica rapa*. *Nature Genetics* 43: 1035–1039.

Wu YC, Rasmussen MD, Bansal MS, Kellis M. 2013. Treefex: Statistically informed gene tree error correction using species trees. *Systematic Biology* 62: 110–120.

Xiang Y, Huang CH, Hu Y, Wen J, Li S, Yi T, Chen H, Xiang J, Ma H. 2017. Evolution of Rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication. *Molecular Biology and Evolution* 34: 262–281.

Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E, Lancet D, Shmueli O. 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* 21: 650–659.

Yang Z. 2007. Paml 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* 24: 1586–1591.

Zeng L, Zhang Q, Sun R, Kong H, Zhang N, Ma H. 2014. Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. *Nature Communications* 5: 4956.

Zeng L, Zhang N, Zhang Q, Endress PK, Huang J, Ma H. 2017. Resolution of deep eudicot phylogeny and their temporal diversification using nuclear genes from transcriptomic and genomic datasets. *New Phytologist* 214: 1358–1354.

Zhang J, Yang JR. 2015. Determinants of the rate of protein sequence evolution. *Nature Reviews Genetics* 16: 409–420.

Zhou Z, Zhou J, Su Z, Gu X. 2014. Asymmetric evolution of human transcription factor regulatory networks. *Molecular Biology and Evolution* 31: 2149–2155.

**Supplementary Material**

The following supplementary material is available online for this article at http://onlinelibrary.wiley.com/doi/10.1111/jse.12712/supplinfo:

**Fig. S1.** Evolutionary rates of orthologous genes are highly correlated in diverse lineages in eudicots. All genes had singleton status (relative to family-level whole-genome duplications (WGDs)) in relative species, and the evolutionary rates were calculated by a comparison of the orthologous pairs in family-level out-lineage species. We reasoned that the rate of evolution (*K*<sub>i</sub>) between the members of each representative pairs would be a good proxy for the rate of evolution of the study genes in a way that is unaffected by the process of duplication (Fig. 1).

**Fig. S2.** Evolutionary rates of orthologous genes are highly correlated in different lineages in Brassicaceae. All genes had singleton status (relative to the *Brassica* rapa whole-genome duplication (WGD)) in relative species, and the evolutionary rates were calculated with a comparison of orthologous pairs.

**Fig. S3.** Evolutionary rates of orthologous genes are highly correlated in different lineages in Fabaceae. All genes have singleton status (relative to the *Glycine max* whole-genome duplication (WGD)) in relative species, and the evolutionary rates were calculated with a comparison of orthologous pairs.

**Fig. S4.** Evolutionary rates of orthologous genes are highly correlated in different lineages in Poaceae. All genes had
singleton status (relative to the Zea mays whole-genome duplication (WGD)) in relative species, and the evolutionary rates were calculated with comparison of orthologous pairs. 

**Fig. S5.** Boxplots showing $K_{\omega}$, $K_s$, and $\omega (K_{\omega}/K_s)$ values for retained clades and singleton clades in family-level out-lineage species. Genes were separated into retained and singleton clades according to their status in the Brassicaceae genome. Values were compared using the Mann–Whitney U-test (paired), and P-values are shown above each pair of boxplots. 

A, Boxplots showing $K_{\omega}$, $K_s$, and $\omega (K_{\omega}/K_s)$ values for retained-clade and singleton-clade genes in the Vitis vinifera and Theobroma cacao gene set (retained clade, $n = 381$; singleton clade, $n = 403$). B, Boxplots showing $K_{\omega}$, $K_s$, and $\omega$ values for retained clades and singleton clade genes in the *Aquilegia coerulea* and *T. cacao* gene set (retained clade, $n = 376$; singleton clade, $n = 391$).

**Fig. S6.** Boxplots showing $K_{\omega}$, $K_s$, and $\omega (K_{\omega}/K_s)$ values for retained and singleton clades in family level out-lineage species. Gene were separated into retained and singleton clades according to their status in the Fabaceae genome. A, Boxplots showing $K_{\omega}$, $K_s$, and $\omega (K_{\omega}/K_s)$ values for retained-clade and singleton-clade genes in the Vitis vinifera and Fragaria vesca gene set (retained clade, $n = 256$; singleton clade, $n = 254$). B, Boxplots showing $K_{\omega}$, $K_s$, and $\omega$ values for retained-clade and singleton-clade genes in the *F. vesca* and *Theobroma cacao* gene set (retained clade, $n = 346$; singleton clade, $n = 347$).

**Fig. S7.** Boxplots showing $K_{\omega}$, $K_s$, and $\omega (K_{\omega}/K_s)$ values for the retained clade (containing more duplicates than the singleton clade) and the singleton clade (containing fewer duplicates than the retained clade, or none) in the Brassicaceae gene set. All genes had singleton status (relative to *Brassica rapa* whole-genome duplication (WGD)) in relative species, and the evolutionary rates were calculated with a comparison of orthologous pairs. The data were separated into A and B genes, according to their status in the *B. rapa* genome. Values were compared using the Mann–Whitney U-test (paired), and P-values are shown above each pair of boxplots.

**Fig. S8.** Boxplots showing $K_{\omega}$, $K_s$, and $\omega (K_{\omega}/K_s)$ values for retained-clade and singleton-clade genes in the Brassicaceae gene set. All genes had singleton status (relative to Glycine max whole-genome duplications (WGDs)) in relative species, and the evolutionary rates were calculated with a comparison of orthologous pairs. The data were separated into A and B genes according to their status in the *G. max* genome. Values were compared using the Mann–Whitney U-test (paired), and P-values are shown above each pair of boxplots.

**Fig. S9.** Boxplots showing $K_{\omega}$, $K_s$, and $\omega (K_{\omega}/K_s)$ values for retained-clade and singleton-clade genes in the Poaceae gene set. All genes had singleton status (relative to Zea mays whole-genome duplications (WGDs)) in relative species, and the evolutionary rates were calculated with a comparison of orthologous pairs. The data were separated into A and B genes according to their status in the *Z. mays* genome. Values were compared using the Mann–Whitney U-test (paired), and P-values are shown above each pair of boxplots.

**Fig. S10.** Boxplots showing the number of genes in the retained clade and the singleton clade in family-level out-lineage species. Numbers of genes greater than 10 are not shown. Values were compared using the Mann–Whitney U-test (unpaired), and P-values are shown above each pair of boxplots. A, Genes were separated into retained clade and singleton clade according to their status in the Brassicaceae genome (retained clade, $n = 912$; singleton clade, $n = 1144$). B, Genes were separated into retained clade and singleton clade according to their status in the Fabaceae genome (retained clade, $n = 809$; singleton clade, $n = 980$).

**Fig. S11.** Boxplots showing the number of genes in the retained clade and the singleton clade in species-level out-lineage species in Brassicaceae. Genes were separated into the retained clade and the singleton clade according to their status in *Brassica rapa*. Values were compared using the Mann–Whitney U-test ($n = 2848$), and P-values are shown above each pair of boxplots.

**Fig. S12.** Boxplots showing the number of genes in the retained and singleton clades in species-level out-lineage species in Fabaceae. Genes were separated into the retained clade and the singleton clade according to their status in Glycine max. Values were compared using the Mann–Whitney U-test ($n = 1350$), and the P-values are shown above each pair of boxplots.

**Fig. S13.** Boxplots showing the number of genes in the retained and singleton clades in species-level out-lineage species in Poaceae. Genes were separated into the retained clade and the singleton clade according to their status in *Zea mays*. Values were compared using the Mann–Whitney U-test ($n = 1281$), and the P-values are shown above each pair of boxplots.

**Fig. S14.** Histograms showing the gene copy number distribution in orthogroups at the level of core eudicots in different species.

**Fig. S15.** Histograms showing the gene copy number distribution in orthogroups at the level of Brassicaceae in different species.

**Fig. S16.** Histograms showing the gene copy number distribution in orthogroups at the level of Fabaceae in different species.

**Fig. S17.** Histograms showing the gene copy number distribution in orthogroups at the level of Poaceae in different species.