OPEN ACCESS Freely available online

PLoS one

# Modes of Gene Duplication Contribute Differently to Genetic Novelty and Redundancy, but Show Parallels across Divergent Angiosperms

**Yupeng Wang[1,2], Xiyin Wang[1,3], Haibao Tang[1,4], Xu Tan[1,4], Stephen P. Ficklin[5], F. Alex Feltus[5,6], Andrew H. Paterson[1,2,4,7,8]***

**1** Plant Genome Mapping Laboratory, University of Georgia, Athens, Georgia, United States of America, **2** Institute of Bioinformatics, University of Georgia, Athens, Georgia, United States of America, **3** College of Life Sciences, Hebei United University, Tangshan, Hebei, China, **4** Department of Plant Biology, University of Georgia, Athens, Georgia, United States of America, **5** Plant and Environmental Sciences, Clemson University, Clemson, South Carolina, United States of America, **6** Department of Genetics and Biochemistry, Clemson University, Clemson, South Carolina, United States of America, **7** Department of Crop and Soil Sciences, University of Georgia, Athens, Georgia, United States of America, **8** Department of Genetics, University of Georgia, Athens, Georgia, United States of America

## Abstract

***Background:*** Both single gene and whole genome duplications (WGD) have recurred in angiosperm evolution. However, the evolutionary effects of different modes of gene duplication, especially regarding their contributions to genetic novelty or redundancy, have been inadequately explored.

***Results:*** In *Arabidopsis thaliana* and *Oryza sativa* (rice), species that deeply sample botanical diversity and for which expression data are available from a wide range of tissues and physiological conditions, we have compared expression divergence between genes duplicated by six different mechanisms (WGD, tandem, proximal, DNA based transposed, retrotransposed and dispersed), and between positional orthologs. Both neo-functionalization and genetic redundancy appear to contribute to retention of duplicate genes. Genes resulting from WGD and tandem duplications diverge slowest in both coding sequences and gene expression, and contribute most to genetic redundancy, while other duplication modes contribute more to evolutionary novelty. WGD duplicates may more frequently be retained due to dosage amplification, while inferred transposon mediated gene duplications tend to reduce gene expression levels. The extent of expression divergence between duplicates is discernibly related to duplication modes, different WGD events, amino acid divergence, and putatively neutral divergence (time), but the contribution of each factor is heterogeneous among duplication modes. Gene loss may retard inter-species expression divergence. Members of different gene families may have non-random patterns of origin that are similar in Arabidopsis and rice, suggesting the action of pan-taxon principles of molecular evolution.

***Conclusion:*** Gene duplication modes differ in contribution to genetic novelty and redundancy, but show some parallels in taxa separated by hundreds of millions of years of evolution.

**Citation:** Wang Y, Wang X, Tang H, Tan X, Ficklin SP, et al. (2011) Modes of Gene Duplication Contribute Differently to Genetic Novelty and Redundancy, but Show Parallels across Divergent Angiosperms. PLoS ONE 6(12): e28150. doi:10.1371/journal.pone.0028150

**Editor:** Stephen R. Proulx, UC Santa Barbara, United States of America

**Received** September 7, 2011; **Accepted** November 2, 2011; **Published** December 2, 2011



**Funding:** AHP appreciates funding from the National Science Foundation (NSF: DBI 0849896, MCB 0821096, MCB 1021718). This study was supported in part by resources and technical expertise from the University of Georgia Research Computing Center, a partnership between the Office of the Vice President for Research and the Office of the Chief Information Officer. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

* E-mail: paterson@plantbio.uga.edu

## Introduction

Whole-genome duplications (WGDs) have occurred in the lineages of plants [1], animals [2,3] and fungi [4,5], with possible consequences including evolution of novel or modified gene functions [6,7,8,9], and/or provision of "buffer capacity" [10,11] or genetic redundancy that increases genetic robustness [12,13,14,15,16,17]. Genome duplication may also increase opportunities for nonreciprocal recombination [18,19,20], permitting or causing duplicated genes to evolve in concert for a period of time. Rapid DNA loss and restructuring of low-copy DNA [21,22,23,24], retrotransposon activation [25,26,27] and epigenetic changes [28,29,30,31,32,33] following WGD may further provide materials for evolutionary change.

Genes may be duplicated by several mechanisms in addition to WGDs, which have been collectively referred to as small scale duplications [34] or single gene duplications [35,36]. Tandem duplicates are consecutive in the genome while proximal duplicates are near one another but separated by a few genes. These two gene duplication modes are presumed to arise through unequal crossing over [36] or localized transposon activities [37]. Dispersed duplicates are neither adjacent to each other in the genome nor within homeologous chromosome segments [38]. Distant single gene transposition may explain the widespread

existence of dispersed duplicates within and among genomes [36]. Distant single gene transposition duplication (referred to as distantly transposed duplication) may occur by DNA based or RNA based mechanisms [35]. DNA transposons such as packmules (rice) [39], helitrons (maize) [40], and CACTA elements (sorghum) [27] may relocate duplicated genes or gene segments to new chromosomal positions (referred to as DNA based transposed duplication). RNA based transposed duplication, often referred to as retrotransposition, typically creates a single-exon retrocopy from a multi-exon parental gene, by reverse transcription of a spliced messenger RNA. It is presumed that the retrocopy duplicates only the transcribed sequence of the parental gene, detached from the parental promoter. The new retrogene is often deposited in a novel chromosomal environment with new (i.e. non-ancestral) neighboring genes and, having lost its native promoter, is only likely to survive as a functional gene if a new promoter is acquired [41,42].

Classical population genetic theory suggests that a likely consequence of gene duplication is reversion to single copy (singleton), unless at least one gene copy evolves new function [8]. More recently, the subfunctionalization model, which proposes that duplicated gene copies might both be retained if they partition the functions of the ancestral gene between them, has described an important modification of the classical model [9,43]. Some studies also show evidence to support the value of genetic redundancy *per se* [10,12,13,14,15,16,17,44,45] or dosage balance [34,46,47,48].

The angiosperms (flowering plants) are an outstanding model in which to elucidate the consequences of gene duplication. All angiosperms are now thought to be paleopolyploids [49], many of which underwent multiple WGDs [50,51]. Traces of past WGDs can often be detected from pairwise syntenic alignments through software such as ColinearScan [52] and multiple alignments using MCScan [53]. Arabidopsis, selected as the first angiosperm genome to be sequenced due to its small genome size and minimal DNA sequence duplication, has experienced two 'recent' WGDs, i.e. since its divergence from other members of the Brassicales clade ($\alpha$ and $\beta$), and a more ancient triplication ($\gamma$) shared with most if not all eudicots [49,51,53]. Likewise, rice appears to have experienced at least two WGDs, one shared with most if not all cereals ($\rho$), and another more ancient event ($\sigma$) [54]. Single gene duplications in angiosperms are also widespread [36,55,56].

One avenue for systematic investigation of functional divergence between duplicate genes is comparison of their spatiotemporal expression profiles, comparing degrees of divergence with proxies of duplication age such as synonymous substitution rates (Ks) between duplicate genes. In Arabidopsis, the rate of protein sequence evolution is asymmetric in >20% of duplicate pairs and functional diversification of surviving duplicate genes has been proposed to be a major feature of the long-term evolution of polyploids [57]. Arabidopsis genes created by large-scale duplication events are more evolutionarily conserved in gene expression than those created by small-scale duplication or those that do not lie in duplicate segments, and the time since duplication is correlated with functional divergence of genes [58]. Further, there may be also a strong positive correlation between expression divergence and non-synonymous mutation (Ka) in Arabidopsis, and the different modes (segmental, tandem and dispersed) of duplication may affect patterns of expression divergence [38]. Arabidopsis duplicated genes show greater expression diversity than singleton genes across closely related species and allopolyploids [59]. In rice, expression correlation is significantly higher for gene pairs from WGDs or tandem duplications than dispersed duplications, and expression divergence is closely related to divergence time [60].

Though many studies have investigated the functional divergence and retention of duplicate genes, conclusions are often contradictory, e.g. gene retention has been attributed to either neofunctionalization [6,7] or genetic redundancy [12,13,14,15,16,17], and expression divergence between duplicate genes has been suggested to be either time dependent [58,60] or selection dependent [38]. The fates of duplicate genes may be influenced by different modes of gene duplication, which have been suggested to retain genes in a biased manner [36]. With much richer expression and annotation data available now than for most prior studies, and improved ability to discern various mechanisms of gene duplication, we find merit in re-examining some existing hypotheses and exploring some new hypotheses regarding the consequences of gene duplication. Here, we related multiple types of genomic data to gene expression divergence in two angiosperm species, Arabidopsis and Oryza (rice), to formally test possible evolutionary patterns (hypotheses). A far richer volume of analyzed microarray data than was available in prior studies improves the robustness of statistical analyses.

## Results

A total of 4,566 Affymetrix Arabidopsis Genome ATH1 Arrays and 508 Affymetrix GeneChip Rice Genome Arrays were used to generate the expression profiles of 22,810 Arabidopsis genes and 27,910 rice genes. We classified gene duplications into six modes: WGD, tandem, proximal, DNA based transposed, retrotransposed and dispersed duplication, according to the procedure shown in Figure 1 and described in methods. Note that in this study, a gene may have up to five potential duplication relationships, depending on the number of BLASTP hits. For WGD duplicates, redundant duplication relationships were removed using co-linearity restrictions. If a gene was created by single gene duplications, all possible duplication relationships were considered. However, redundant duplication relationships in single gene duplications did not enlarge the gene set created by each duplication mode. In a distantly transposed duplication, one duplicate gene is the parental (ancestral) copy while the other is the transposed (derived) copy, at a novel locus. Dispersed duplications, which we cannot attribute to specific mechanisms, are regarded as a control group. The number of pairs of duplicate genes and number of unique genes (i.e. number of created genes) in each mode of duplication is summarized in Table 1. A total of 2,981 $\alpha$, 1,161 $\beta$ and 417 $\gamma$ WGD duplicate pairs in Arabidopsis; and 1,712 $\rho$ and 568 $\gamma$ WGD duplicate pairs in rice, have expression profiles. In this study, the degree of similarity between the expression profiles of a pair of genes across all experiments is measured by the Pearson's correlation coefficient ($r$). To express in positive values the evolution of gene expression between duplicates or orthologs, we use the term "expression divergence", measured by $1-r$ [61,62].

### Gene duplication modes contribute differentially to genetic novelty and redundancy

Expression divergence between duplicate genes was compared across modes of duplication (Figure 2). The trends of expression divergence between duplicates in Arabidopsis and rice are very similar: DNA based transposed duplication≈retrotransposed duplication > dispersed duplication > proximal duplication > WGD≈tandem duplication (both ANOVA model involving all duplication modes and Tukey's HSD test between adjacent duplication modes are significant at $\alpha = 0.05$). Although retrotransposed duplications have a little higher average expression divergence than DNA based transposed duplications, the difference is not significant ($P$-value>0.05). WGDs result in a little higher

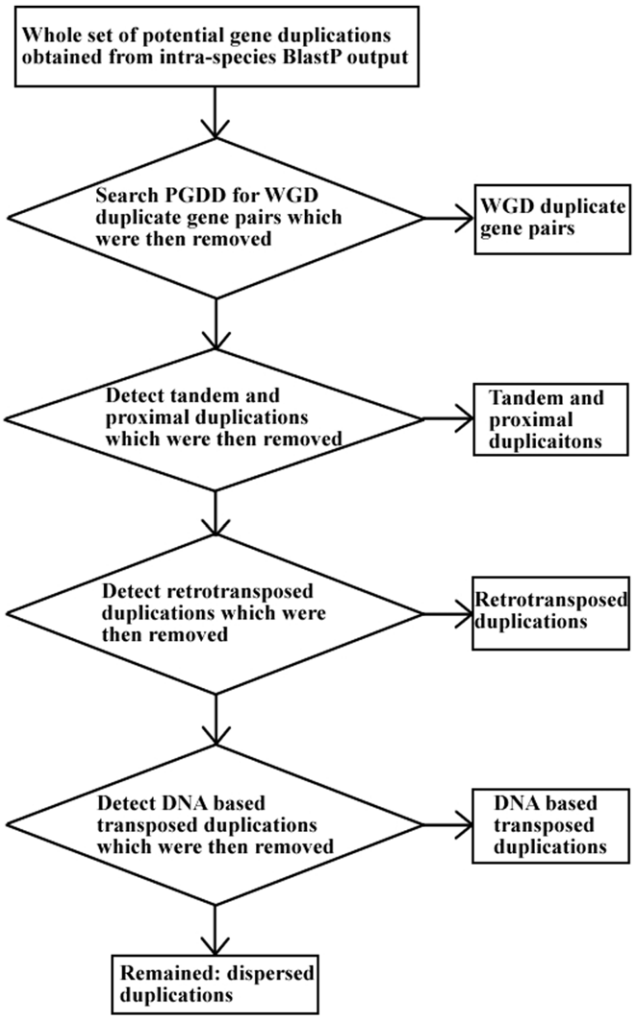


**Figure 1. Flowchart of the procedure for classifying gene pairs based on mode of duplication.**
doi:10.1371/journal.pone.0028150.g001

expression divergence than tandem duplications in Arabidopsis but the difference is not significant in rice.

Despite the relatively fast evolution of gene expression shown by distantly transposed duplications, a tendency toward co-expression between genes duplicated by all modes can be observed by comparison with 10,000 randomly selected gene pairs (Figure 2). Furthermore, we used $r<0.371$ and $r<0.621$ (95% quantile of the $r$ values obtained from random gene pairs) as criteria for determining that two duplicate genes have diverged in expression in Arabidopsis and rice respectively [57,63]. The proportions of divergent expression between genes duplicated by different modes are shown in Table 2. All these data suggest that the extent of expression divergence of retained duplicates is affected by the duplication mechanism: WGD and tandem duplicates are more likely to maintain their original expression patterns, proximal duplications show intermediate divergence, and distantly transposed duplications tend to have the biggest changes of gene expression profiles.

Computationally, genetic redundancy may be inferred from simultaneous conservation in protein sequences that determine molecular functions, and expression patterns which determine biological processes [64,65]. WGD and tandem duplicates tend to be simultaneously conserved in protein sequences (using 25% quartile of Ka of all duplicate pairs, i.e. $<0.329$ in Arabidopsis and $<0.383$ in rice, as criteria) and in gene expression (using $r \geq 0.371$ in Arabidopsis and $r \geq 0.621$ in rice as criteria), while distantly transposed and dispersed duplicates have a random association (assuming that conservation in protein sequences and gene expression were independent in the pooled duplicate genes) between these parameters, and proximal duplicates fall in between (Table 3).

Expression levels differ between the genes created by different duplication modes (Figure 3). WGD and dispersed duplicates have higher gene expression levels than tandem, proximal and distantly transposed duplications (2-sample $t$-tests are significant at $\alpha = 0.05$). The higher expression of WGD duplicates is consistent with their retention due to dosage amplification, a theory which has been proven in yeast [47,66,67]. Potentially transposon mediated gene duplications including tandem, proximal and distantly transposed duplications tend to be associated with lower gene expression levels than other duplication modes (Figure 3). Dispersed duplication, with unclear genetic mechanisms so far, is associated with gene expression levels comparable to WGD.

## Expression divergence following polyploidy

Since its divergence from other Brassicales, Arabidopsis experienced two WGDs ($\alpha$ and $\beta$), while sharing a more ancient genome triplication ($\gamma$) with all rosids and perhaps all eudicots [49,51,53]. Rice has experienced two WGDs: the $\rho$ event shared with all Poaceae, and the more ancient $\sigma$ event [54]. Although expression divergence has been compared between WGD and single gene duplications [38,58,60], the combinational effects of different WGD events on expression divergence have not been

**Table 1.** Numbers of pairs of duplicate genes and unique genes in each mode of gene duplication.

| Mode of duplication | Number of pairs of duplicate genes (number of those having complete expression profiles) | | Number of unique genes (number of those having expression profiles) | |
|---|---|---|---|---|
| | Arabidopsis | Rice | Arabidopsis | Rice |
| WGD | 6,572 (4,979) | 3,593 (2,530) | 9,455 (8,089) | 5,723 (4,829) |
| Tandem | 2,055 (1,055) | 1,741(947) | 1,586 (977) | 2,948 (2,116) |
| Proximal | 3,113 (1,456) | 3,816 (1,990) | 669 (379) | 1,038 (714) |
| DNA based transposed | 6,367 (4,088) | 8,061 (5,225) | 2,230 (1,572) | 2,948 (2,116) |
| Retro- transposed | 497 (300) | 940 (681) | 271 (1,71) | 491 (391) |
| Dispersed | 34,887 (26,127) | 30,574 (21,385) | 7,411 (6,182) | 8,313 (6,960) |

doi:10.1371/journal.pone.0028150.t001

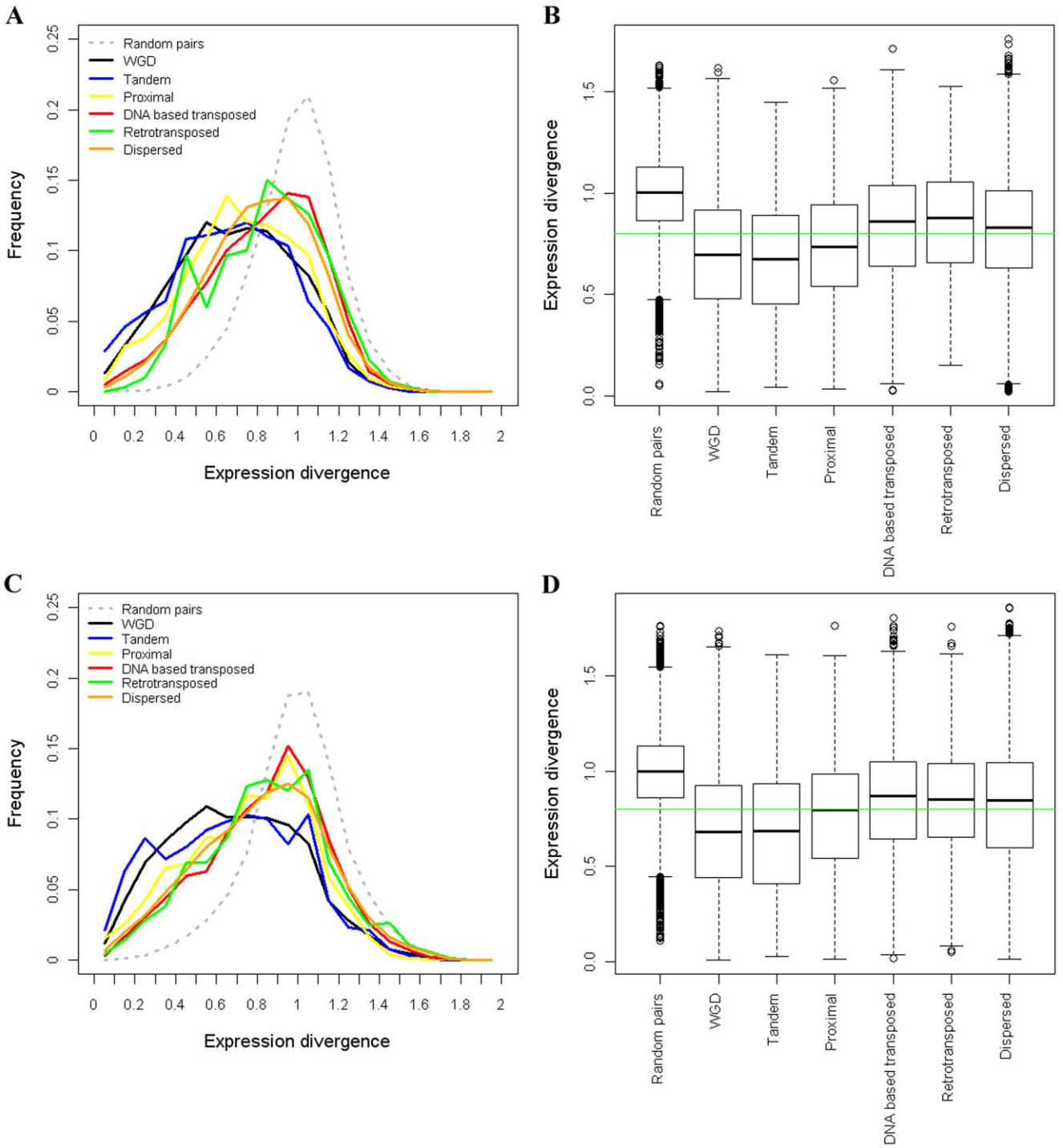


**Figure 2. Comparison of expression divergence among different modes of gene duplication.** (A) Comparison of distributions of expression divergence in Arabidopsis. (B) Comparison of levels of expression divergence in Arabidopsis. (C) Comparison of distributions of expression divergence in rice. (D) Comparison of levels of expression divergence in rice. Green lines in (B, D) indicate average expression divergence across duplication modes.
doi:10.1371/journal.pone.0028150.g002

addressed. We propose that WGD events themselves, together with the subsequent 'adaptation' of the resulting genome to the newly-duplicated state, may accelerate evolution, contributing to variation in expression divergence sometimes attributed to time (usually measured by Ks) alone [58,60].

To further investigate the combinational effects of multiple WGD events, we compared the expression divergence of duplicates from different WGD events (Figure 4). Not surprisingly, expression divergence between the WGD duplicates of more ancient events tends to be larger: $\gamma$ duplicates $>$ $\beta$ duplicates $>$ $\alpha$ duplicates in Arabidopsis, and $\sigma$ duplicates $>$ $\rho$ duplicates in rice (both ANOVA model involving all WGD events and Tukey's HSD test between adjacent WGD events are significant at $\alpha = 0.05$). Next, we fitted a curve between expression divergence

**Table 2.** Proportion of divergent gene expression between duplicates in each mode of gene duplication.

| Species | WGD | Tandem duplication | Proximal duplication | DNA based transposed duplication | Retrotransposed duplication | Dispersed duplication |
|---|---|---|---|---|---|---|
| Arabidopsis | 0.577 | 0.555 | 0.644 | 0.759 | 0.767 | 0.759 |
| Rice | 0.813 | 0.780 | 0.865 | 0.916 | 0.921 | 0.904 |

doi:10.1371/journal.pone.0028150.t002

and Ks for each WGD event using a smooth spline with 10 degrees of freedom available in R packages (Figure 4). We found no significant correlation between expression divergence and Ks within the more ancient Arabidopsis β duplicates ($r = 0.036$, $P$-value = 0.241) or γ duplicates ($r = -0.008$, $P$-value = 0.883), or rice σ duplicates ($r = 0.045$, $P$-value = 0.307) but correlations are significant within the most recent Arabidopsis α duplicates ($r = 0.126$, $P\text{-value} = 1.364 \times 10^{-11}$) and rice ρ duplicates ($r = 0.105$, $P\text{-value} = 2.054 \times 10^{-5}$). Further, we conducted a power analysis for these correlations. We found that at $\alpha = 0.05$, the non-significant correlations (β, γ and σ duplicates) did not have higher power than conventionally desired (>0.8) while significant correlations (α and ρ duplicates) had power greater than 0.98, confirming that the relationship between expression divergence and Ks differs among different WGD events.

WGD events themselves influence gene expression divergence, with more ancient WGD duplicated genes likely to have greater expression divergence than more recent duplications, even if both have similar Ks (Figure 5). To support this hypothesis statistically, we coded the α, β and γ events by 1, 2 and 3 in Arabidopsis and the ρ and σ events by 1 and 2 in rice. Then different linear regression models of expression divergence on Ks and/or WGD codes were fit in Arabidopsis and rice respectively. All regression models and their coefficients were statistically significant. For both Arabidopsis and rice, the model which counts both Ks and the number of WGD events that duplicate genes underwent results in the highest adjusted $R^2$ and lowest Akaike information criterion (AIC) (Table 4) with significant nonzero slopes of all coefficients, supporting the hypothesis that WGD events themselves, in addition to Ks, can lead to increased expression divergence between duplicates.

Selection after WGD events may constrain expression divergence of some duplicates. To examine this question, we studied the 25% of WGD duplicate pairs with most conserved expression at each WGD event. At a $P$-value threshold of 0.05 by Fisher's exact test (corrected for multiple tests), specific GO terms/Pfam domains were associated with conserved expression at each WGD event, and some recurred across different WGD events, e.g. transcription factor activity (GO:0003700) and ribosome (GO:0005840) for Arabidopsis α and γ and rice ρ events; protein biosynthesis (GO:0006412) for Arabidopsis α and β and rice ρ events (Table S1). In contrast, WGD duplicates with divergent expression (25% of pairs with highest $d$ values at each event) showed little or no enrichment of specific GO terms/Pfam domains and functional terms did not recur between different WGD events.

## Expression divergence between Arabidopsis and rice

In that most angiosperms share most genes, changes in expression may be fundamental to angiosperm biodiversity. Previous studies have associated duplicated genes with greater expression diversity than singletons in closely related species of both animals [68] and plants [59]. However, it has been difficult to extend such comparisons to more distant species such as Arabidopsis, a eudicot, and rice, a monocot, due to greater difficulty discerning orthology or paralogy. To facilitate the comparison of gene expression data generated by different microarray platforms, we adopted a conceptual framework of comparing co-expression patterns across species [69] (see Methods). Further, we restricted our study to 2,012 gene pairs suggested both by DNA sequence similarity and by synteny/collinearity to be orthologs between Arabidopsis and rice, downloaded from the PGDD database [51,53]. The comparison of expression divergence between different types of orthologs shows the following trend: duplicate-duplicate>singleton-duplicate>singleton-singleton (Figure 6), with $P$-values of 0.049 between duplicate-duplicate and singleton-duplicate and 0.010 between singleton-duplicate and singleton-singleton using two-sample $t$-tests. This finding supports that singletons are more conserved in expression than duplicated genes, consistent with the hypothesis that one consequence of gene duplication is increased expression diversity.

## Expression divergence may be correlated with both Ks and Ka

Divergence in coding sequences can be denoted by Ks, which indicates putatively-neutral mutations that are synonymous at the amino acid level, or by Ka, which indicates altered amino acids suggestive of the action of selection on gene function. The correlations between expression divergence and coding sequence divergence in angiosperms have been widely discussed [38,58,60] but conclusions were inconsistent: Casneuf et al. and Li et al.

**Table 3.** Proportion of conservation in both protein sequences and gene expression between duplicates in each mode of gene duplication.

| Species | WGD | Tandem duplication | Proximal duplication | DNA based transposed duplication | Retro-transposed duplication | Dispersed duplication | Expected |
|---|---|---|---|---|---|---|---|
| Arabidopsis | 0.335 | 0.328 | 0.231 | 0.071 | 0.051 | 0.038 | 0.071 |
| Rice | 0.140 | 0.170 | 0.099 | 0.027 | 0.023 | 0.021 | 0.041 |

doi:10.1371/journal.pone.0028150.t003

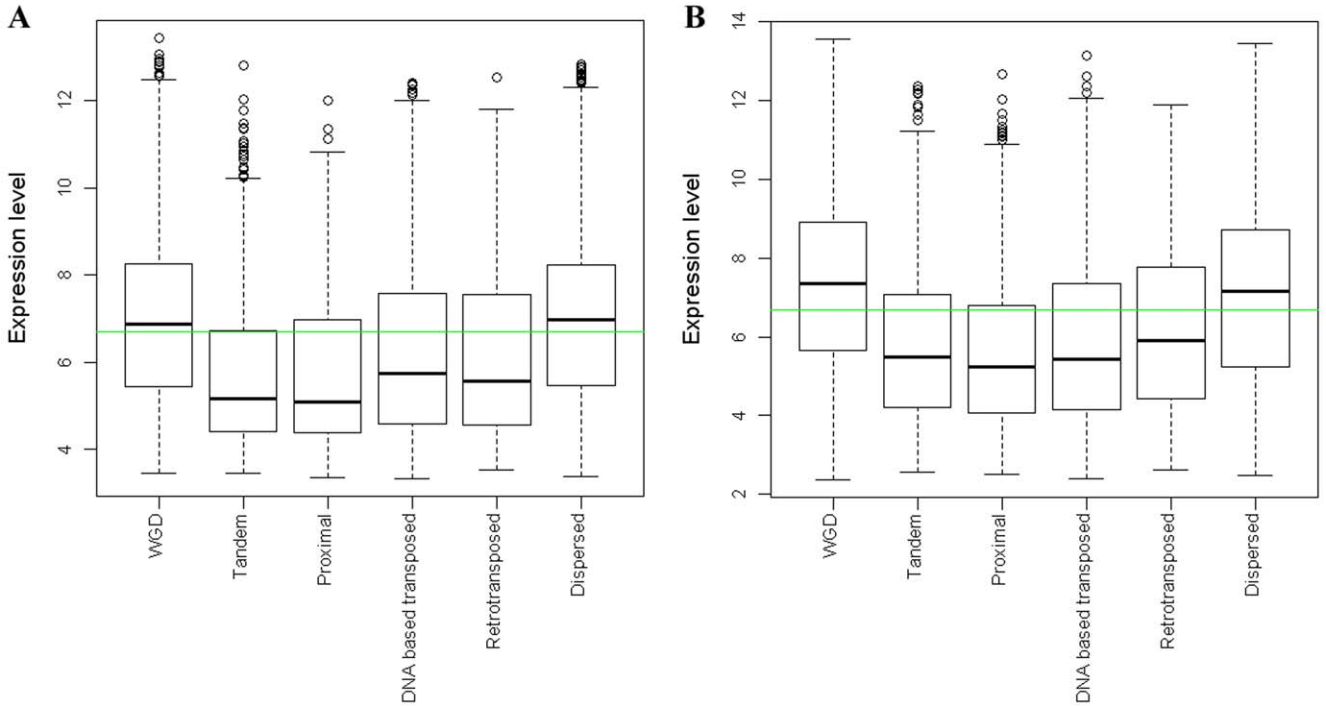


**Figure 3. Comparison of expression levels between genes created by different duplication modes.** (A) Comparison of expression levels between Arabidopsis genes created by different duplication modes. (B) Comparison of expression levels between rice genes created by different duplication modes. Green lines indicate average expression levels.
doi:10.1371/journal.pone.0028150.g003

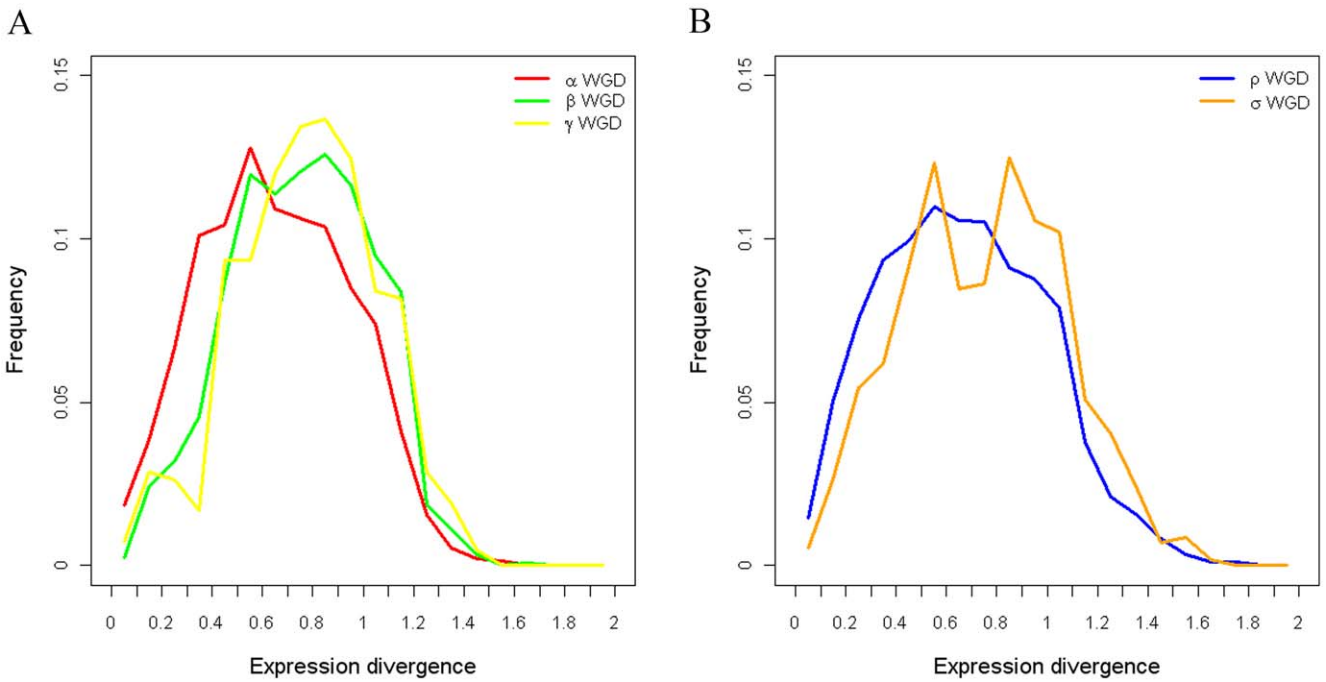


**Figure 4. Comparison of distributions of expression divergence among different WGD events.** (A) Comparison of distributions of expression divergence among different Arabidopsis WGD events. (B) Comparison of distributions of expression divergence among different rice WGD events. α, β and ρ were relatively recent WGD events, while γ and σ were more ancient WGD events.
doi:10.1371/journal.pone.0028150.g004

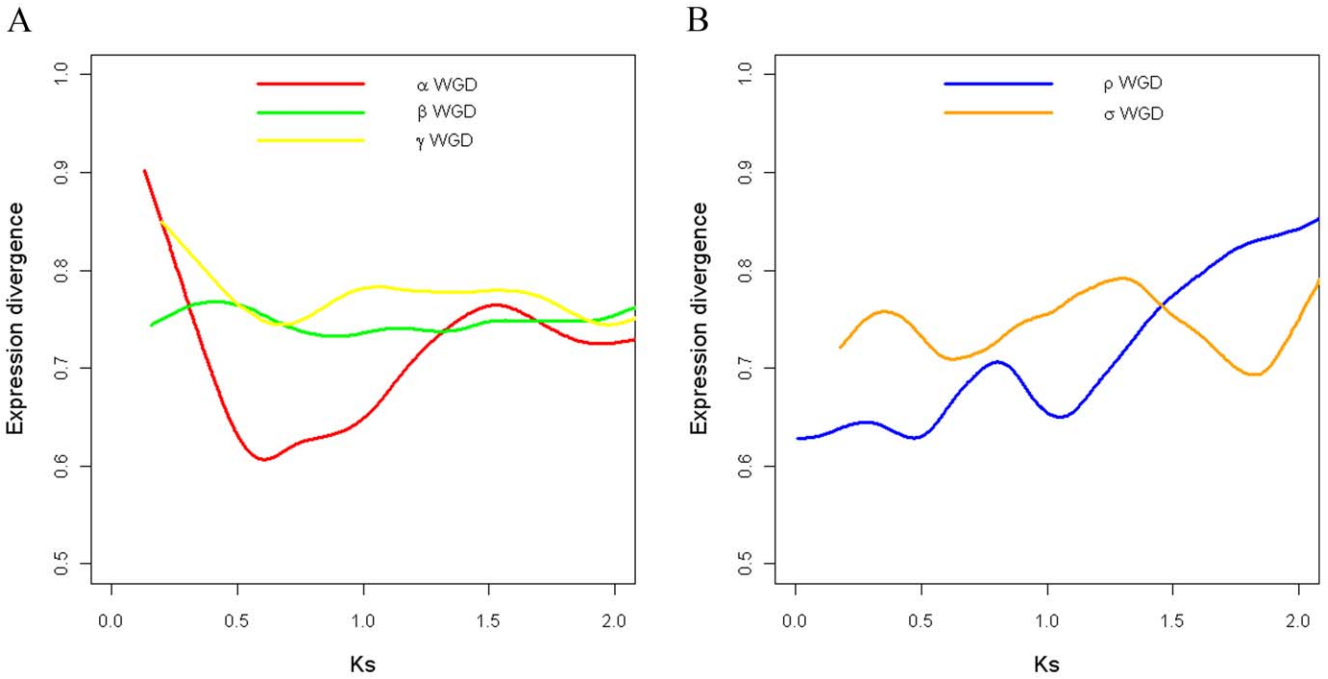


**Figure 5. Fitted smooth spline curves between expression divergence and Ks for different WGD events.** (A) Fitted smooth spline curves between expression divergence and Ks for different Arabidopsis WGD events. (B) Fitted smooth spline curves between expression divergence and Ks for different rice WGD events. α, β and ρ were relatively recent WGD events, while γ and σ were more ancient WGD events.
doi:10.1371/journal.pone.0028150.g005

suggested that Ks is closely correlated with gene expression divergence, while Ganko et al. found little correlation. Since microarray data contain a high level of noise and previous studies often relied on small sets of microarray data or only one species, our analysis of "all arrays" and two highly-divergent species may have broader inference space.

The distributions of Ka or Ks differ markedly for different gene duplication modes, but are relatively consistent in Arabidopsis and rice (Figure 7). Tandem/proximal and WGD duplicates have qualitatively lower Ks (putatively reflecting younger age) than distantly transposed (DNA and RNA) or dispersed duplicates, the distinction being much clearer in the small genome of Arabidopsis (Figure 7A) than the 3× larger and more repeat-rich genome of rice (Figure 7B). Within these qualitative distinctions, quantitative differences among the categories are also evident and largely consistent, with relative Ks (putatively age) of duplications following the trend of: dispersed > distantly transposed > WGD > proximal > tandem (both ANOVA model involving all duplication modes and Tukey's HSD test between adjacent duplication modes are significant at $\alpha = 0.05$). Retrotransposed duplicates differ slightly in the two taxa, being similar to DNA based transposed duplicates in Arabidopsis, and to dispersed duplicates in rice. The trend of Ka shows the same qualitative distinction as that of Ks (Figure 7C and 7D), but differing in the quantitative trend with amino-acid altering mutation frequencies being retrotransposed > dispersed > DNA based transposed > proximal≈WGD≈tandem (both ANOVA model involving all duplication modes and Tukey's HSD test between adjacent duplication modes are significant at $\alpha = 0.05$). WGD duplicates are more functionally constrained, with higher Ks but equal or lower Ka than proximal duplicates. These data do not show the conventional L-shaped distribution for dispersed and distantly

**Table 4.** Linear regression of expression divergence (*d*) on Ks and WGD events (*W*).

| Regression model | Coefficient (*P*-value) | | | Adjusted $R^2$ | AIC |
|---|---|---|---|---|---|
| | $a$ | $b_1$ | $b_2$ | | |
| Arabidopsis | | | | | |
| $d = a + b_1 \cdot Ks$ | 0.593 ($<2.2\times10^{-16}$) | 0.079 ($<2.2\times10^{-16}$) | - | 0.027 | −10706.164 |
| $d = a + b_2 \cdot W$ | 0.577 ($<2.2\times10^{-16}$) | - | 0.074 ($<2.2\times10^{-16}$) | 0.027 | −10706.330 |
| $d = a + b_1 \cdot Ks + b_2 \cdot W$ | 0.559 ($<2.2\times10^{-16}$) | 0.050 ($1.15\times10^{-8}$) | 0.047 ($1.05\times10^{-8}$) | 0.034 | −10736.930 |
| Rice | | | | | |
| $d = a + b_1 \cdot Ks$ | 0.624 ($<2.2\times10^{-16}$) | 0.081 ($1.84\times10^{-7}$) | - | 0.012 | −4913.4477 |
| $d = a + b_2 \cdot W$ | 0.587 ($<2.2\times10^{-16}$) | - | 0.079 ($8.28\times10^{-7}$) | 0.011 | −4916.3561 |
| $d = a + b_1 \cdot Ks + b_2 \cdot W$ | 0.557 ($<2.2\times10^{-16}$) | 0.063 ($1.44\times10^{-4}$) | 0.058 ($6.82\times10^{-4}$) | 0.017 | −4925.9138 |

doi:10.1371/journal.pone.0028150.t004

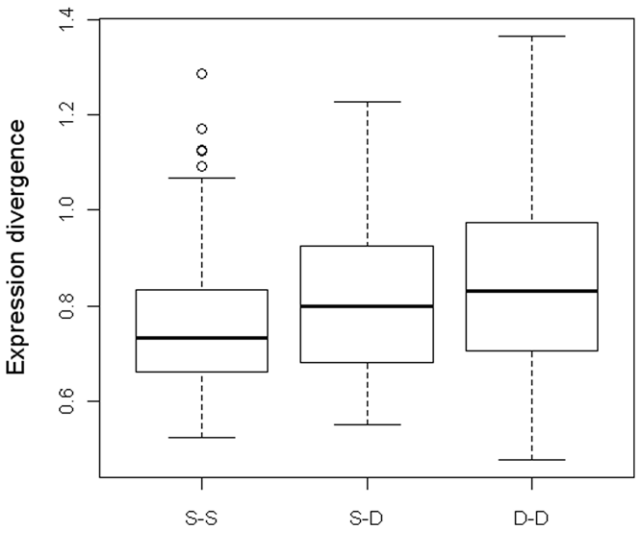


**Figure 6. Comparison of expression divergence between different types of Arabidopsis-rice orthologs: singleton-singleton (S-S), singleton-duplicate (S-D) and duplicate-duplicate (D-D).**
doi:10.1371/journal.pone.0028150.g006

transposed duplicates, because the filters employed in gene selection focus this analysis only on genes that have survived a long time, implying that the genes serve important functions.

Relationships between coding sequence divergence and expression divergence are heterogeneous, and differ among gene duplication modes. For WGD duplicates, expression divergence is significantly correlated with both Ka and Ks in both Arabidopsis and rice, although the strength of the correlations is progressively weaker for more ancient duplications and in some cases reaches non-significance (Table 5). Expression divergence is also significantly correlated with both Ka and Ks among proximal duplicates. Tandem duplicates differ in the two taxa, with those of rice resembling WGD genes with expression divergence significantly correlated with both Ka and Ks, and those of Arabidopsis resembling distantly transposed duplications with marginal and sometimes non-significant correlation.

While age and functional divergence are more closely related to expression divergence in WGD genes than those resulting from other duplication modes, this does not reflect a lack of expression divergence among other gene duplicates. Indeed, proximal duplication is associated with higher expression divergence than WGD, despite its smaller average Ks. Likewise, DNA based transposed duplication is associated with higher expression divergence than dispersed duplication, despite smaller Ks (Table 6).

In partial summary, expression divergence between duplicate genes may be affected by duplication modes, as well as by the 'age' (Ks) of the duplicated genes, i.e. gene expression divergence may differ among duplication modes at the same Ks or Ka levels. To further validate this claim, we fit a smooth spline curve between expression divergence and Ks or Ka for each duplication mode (Figure 8). While these curves fluctuate markedly, at fixed Ks or Ka levels distantly transposed duplications (for example) are generally associated with higher expression divergence between duplicates than WGD or tandem duplications.

## DNA methylation of the promoter regions has little impact on expression divergence

Epigenetic mechanisms such as DNA methylation have been suggested to potentially differentiate newly arisen duplicate genes [32,70] as well as orthologous genes across closely related species [59]. Transcriptional silencing has often been associated with DNA methylation in promoter regions [71,72]. Using data on genome-wide DNA methylation status for both Arabidopsis and rice [73], we examined whether DNA methylation status in promoter regions is related to expression divergence between duplicates or between orthologs. This comparison carries an inherent assumption that methylation patterns are relatively static and generally apply to all of the microarray studies. A gene promoter region was considered to be methylated if two or more adjacent probes are methylated within the region [72]. Proportions of pairs of duplicates that differ in DNA methylation status in promoter regions, separated by gene duplication modes, are summarized in Table 7. Distantly transposed duplications appear somewhat more likely to differ in DNA methylation status than other duplication modes. However, the duplicate genes that differ in DNA methylation status in promoter regions do not have more divergent expression than those that have the same DNA methylation status, within any duplication mode (negative data are not shown). Likewise, different methylation status among orthologs also showed no significant relationship to expression divergence, although we confirmed that singletons are a little more likely to be methylated in promoter regions than duplicates (Table 8), as proposed by others [59]. These analyses suggest that the mechanisms by which DNA methylation status affects expression divergence between homologous genes may be complicated, and direct association may not be informative for unraveling such mechanisms.

## Gene family members may have non-random patterns of origin

The diversity of gene duplication mechanisms and patterns of gene expression divergence raise questions about how gene families expand and how their members have been retained in the history of evolution. WGD duplicates are differentially retained across different gene functional classifications [10,34, 57,74]. However, we suggest that gene families may be more informative units than functional terms for investigating patterns of gene origin, as duplication relationships in gene families are clearer. Based on our findings above, both functional divergence and redundancy may contribute to retention of duplicate genes. Furthermore, because the degrees of functional diversification are not equal across gene families and gene duplication modes add additional heterogeneity to patterns of functional divergence, it is possible that gene family members may have non-random patterns of origin, e.g. the gene families with high functional diversification may be enriched with distantly transposed duplications while those families contributing to genetic redundancy are likely to be enriched with WGD duplications.

To examine these questions, we investigated the gene duplication modes of 126 Arabidopsis and 24 rice published gene families of 10 or more genes, available at TAIR (http://www.arabidopsis.org/) and Michigan State University (http://rice.plantbiology.msu.edu/) respectively. By using Bonferroni-corrected Fisher's exact test, we found that 64 (50.8%) Arabidopsis gene families and 19 (79.2%) rice gene families are enriched for at least one gene duplication mode at $\alpha = 0.05$ (Table S2). For example, DNA based transposed duplications are enriched in disease resistance gene homologs and the cytochrome P450 gene family (Figure 9 A–C). Disease resistance gene homologs, most of which have nucleotide binding site-leucine rich repeat (NBS-LRR) domains, express at different levels and tissue specificities, and function in diverse biological processes in Arabidopsis [75]. P450s

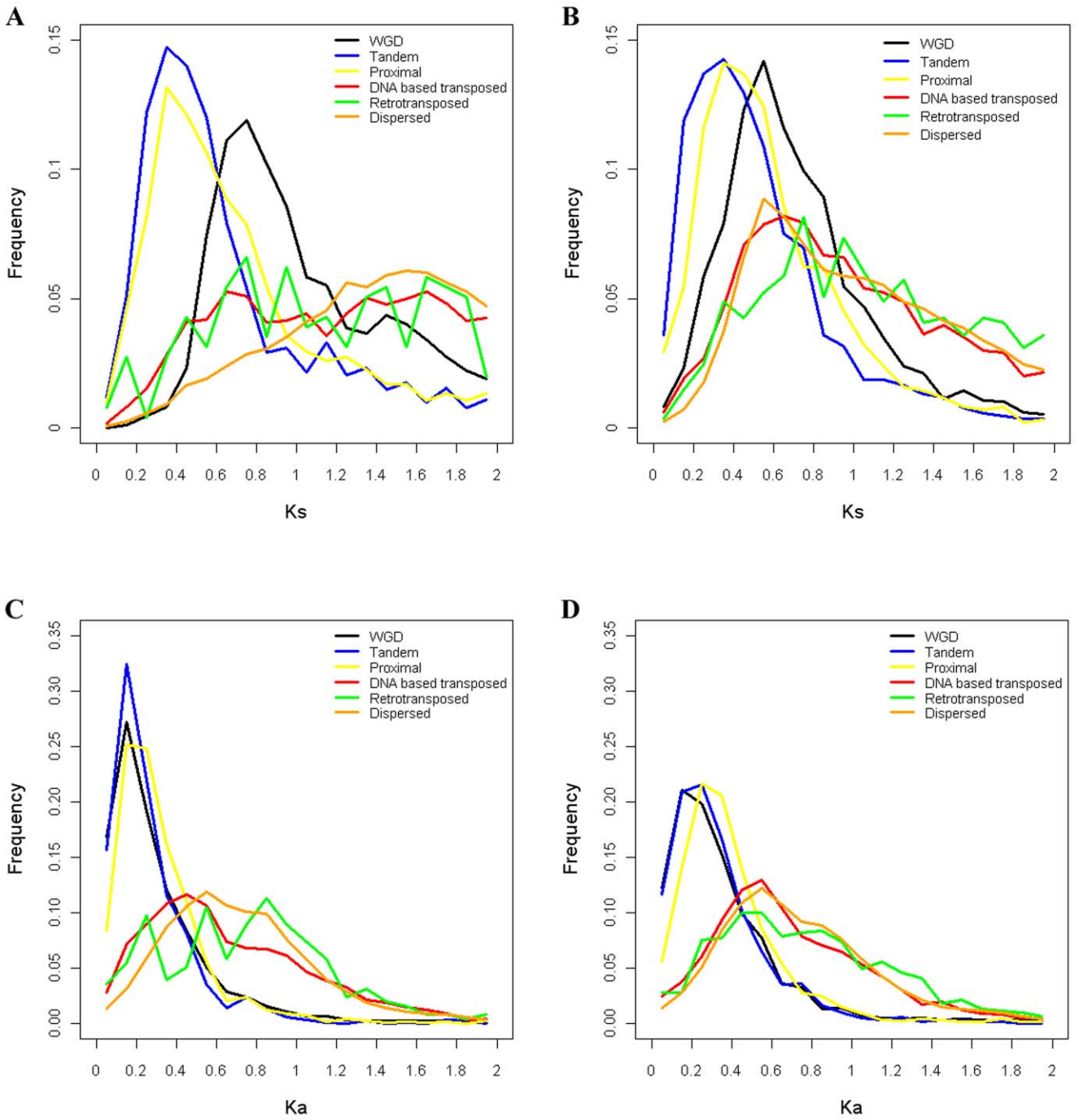


**Figure 7. Comparison of Ks and Ka distributions for gene pairs duplicated by different modes.** (A) Comparison of Ks distributions in Arabidopsis. (B) Comparison of Ks distributions in rice. (C) Comparison of Ka distributions in Arabidopsis. (D) Comparison of Ka distributions in rice. doi:10.1371/journal.pone.0028150.g007

also express in many tissues in a tissue specific manner and are involved in diverse metabolic processes [76,77]. The cytochrome P450 family also shows enrichment for DNA based transposed duplications in rice. Thus, these two gene families may have achieved functional and expression diversity through some combination of transposition activity and retention of distantly transposed duplicates. Interestingly, these two families are also enriched with proximal duplications, again often associated with greater expression divergence than WGD despite generally similar coding sequence divergence.

WGD duplicates are enriched in other gene families, such as the cytoplasmic ribosomal protein gene family, and C2H2 zinc finger proteins (Figure 9 D–F). In Arabidopsis, a large number of ribosomal genes are co-regulated [78]. C2H2 zinc finger proteins have been shown to be involved in some basic biological processes such as transcriptional regulation, RNA metabolism and chromatin-remodeling [79]. Furthermore, C2H2 zinc finger proteins are enriched with retained WGD duplicates in both Arabidopsis and rice. Our analyses suggest that gene family members may have common non-random patterns of origin, that recur independently

**Table 5.** Correlations between expression divergence (*d*) and coding sequence divergence.

| Types of homologs | Number of valid gene pairs | Pearson correlation (*P*-value) between *d* and | |
|---|---|---|---|
| | | Ka | Ks |
| Arabidopsis duplicates | | | |
| WGD | 4,682 | 0.238 ($<2.2\times10^{-16}$) | 0.176 ($<2.2\times10^{-16}$) |
| α | 2,858 | 0.247 ($<2.2\times10^{-16}$) | 0.126 ($1.364\times10^{-11}$) |
| β | 1,068 | 0.146 ($1.791\times10^{-6}$) | 0.036 (0.241) |
| γ | 371 | 0.060 (0.253) | −0.008 (0.883) |
| Tandem | 1,033 | 0.015 (0.635) | 0.115 ($2.137\times10^{-4}$) |
| Proximal | 1,426 | 0.057 (0.032) | 0.113 ($1.891\times10^{-5}$) |
| DNA based transposed | 3,662 | 0.052 (0.002) | 0.023 (0.173) |
| Retrotransposed | 257 | 0.042 (0.504) | 0.142 (0.023) |
| Dispersed | 23,360 | 0.046 ($3.243\times10^{-12}$) | 0.047 ($1.087\times10^{-12}$) |
| Rice duplicates | | | |
| WGD | 2,390 | 0.112 ($4.006\times10^{-8}$) | 0.112 ($3.984\times10^{-8}$) |
| ρ | 1,630 | 0.099 ($6.519\times10^{-5}$) | 0.105 ($2.054\times10^{-5}$) |
| σ | 521 | 0.059 (0.177) | 0.045 (0.307) |
| Tandem | 919 | 0.091 (0.006) | 0.087 (0.008) |
| Proximal | 1,898 | 0.084 ($2.389\times10^{-4}$) | 0.095 ($3.604\times10^{-5}$) |
| DNA based transposed | 4,687 | 0.056 ($1.126\times10^{-4}$) | 0.017 (0.255) |
| Retrotransposed | 613 | 0.008 (0.839) | 0.037 (0.361) |
| Dispersed | 19,397 | 0.037 ($2.225\times10^{-7}$) | 0.017 (0.021) |
| Arabidopsis-rice orthologs | 1,290 | 0.108 ($9.468\times10^{-5}$) | 0.003 (0.901) |

doi:10.1371/journal.pone.0028150.t005

in different evolutionary lineages (such as monocots, and dicots, studied here), and that such patterns may result from specific biological functions and evolutionary needs.

## Discussion

In two species that sample a wide range of tissues and physiological conditions in major angiosperm lineages diverged by about 140–170 million years [80] and affected by at least 5 different genome duplication events, we have compared expression divergence between positional orthologs and between genes duplicated by several additional mechanisms. Both neo-functionalization and genetic redundancy can result in retention of duplicate genes. WGD duplicates generally are more frequently associated with genetic redundancy than genes resulting from other duplication modes, partly due to dosage amplification. Tandem duplications also contribute to genetic redundancy, while other duplication modes are more frequently associated with evolutionary novelty. Potentially transposon mediated gene duplications tend to reduce gene expression levels. Expression divergence between duplicates is discernibly related to duplication modes, WGD events, Ka, Ks, and possibly the DNA methylation status of their promoter regions. However, the contribution of each factor is heterogeneous among duplication modes, and new factors as well as combinatorial effects of different factors are worth further investigation. Gene loss may retard inter-species expression divergence, as singletons are generally more conserved in gene expression than duplicates. Members of different gene families have non-random patterns of origin, and such patterns may be similar between Arabidopsis and rice.

The use of large volumes of data and inclusion of as many genes as possible may help to mitigate factors specific to particular developmental states, noise associated with microarray data, and bias reflecting features specific to particular gene families. For example, we have found that the correlations between expression divergence and Ks are not consistent within gene duplication

**Table 6.** Comparisons of expression divergence and Ks between WGD and proximal duplication, and between dispersed and DNA based transposed duplication.

| Duplication modes | Arabidopsis | | Rice | |
|---|---|---|---|---|
| | Mean d (*P*-value by *t*-test) | Mean Ks (*P*-value by *t*-test) | Mean d (*P*-value by *t*-test) | Mean Ks (*P*-value by *t*-test) |
| WGD vs Proximal | 0.690 vs 0.731 ($2.912\times10^{-6}$) | 1.162 vs 0.816 ($<2.2\times10^{-16}$) | 0.690 vs 0.758 ($1.47\times10^{-12}$) | 0.759 vs 0.619 ($<2.2\times10^{-16}$) |
| Dispersed vs DNA based transposed | 0.813 vs 0.825 (0.019) | 1.710 vs 1.490 ($<2.2\times10^{-16}$) | 0.821 vs 0.825 (0.490) | 1.169 vs 1.490 ($<2.2\times10^{-16}$) |

doi:10.1371/journal.pone.0028150.t006

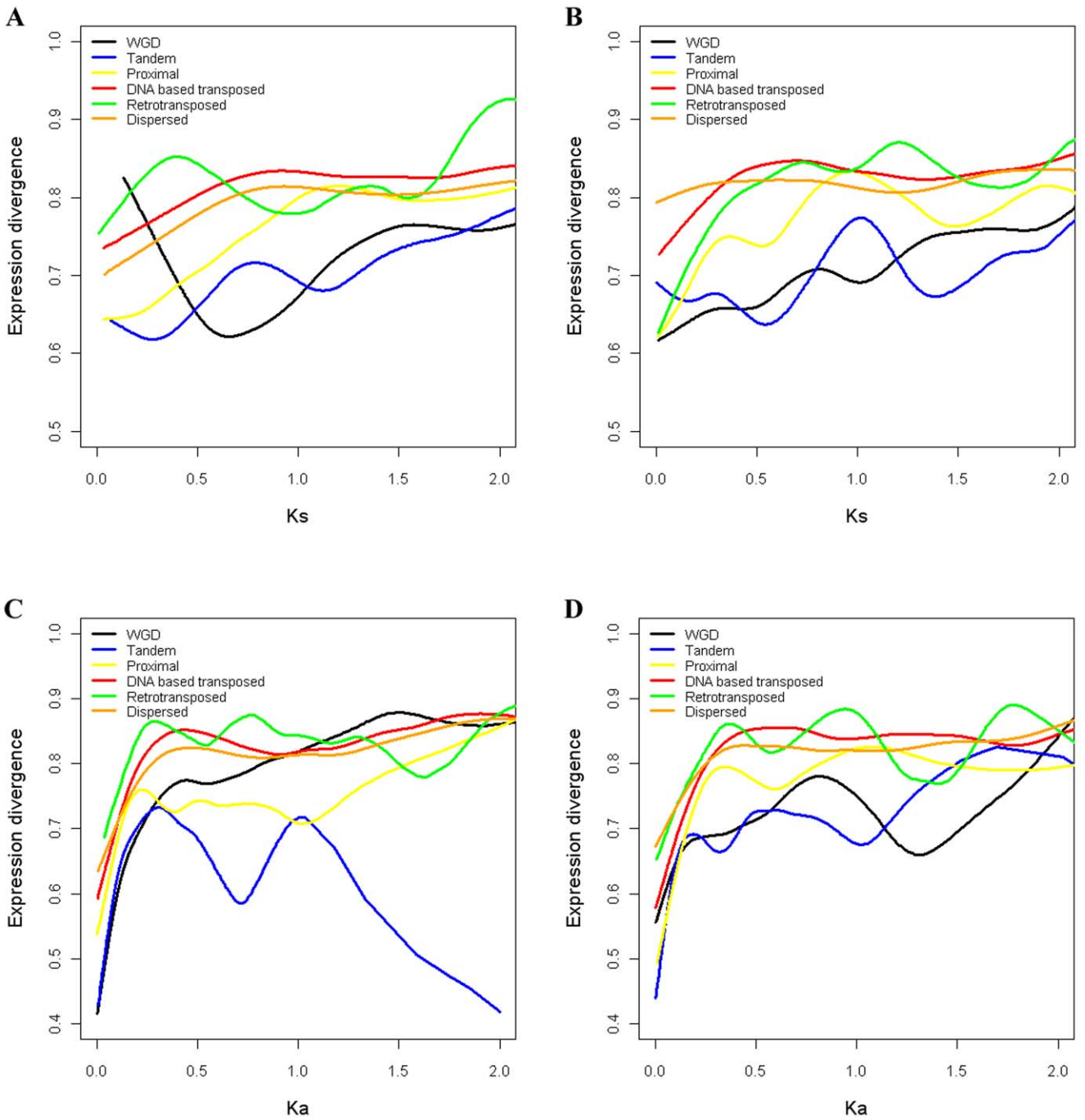


**Figure 8. Fitted smooth spline curves between expression divergence and Ks or Ka for different modes of gene duplication.** (A) Fitted smooth spline curves between expression divergence and Ks in Arabidopsis. (B) Fitted smooth spline curves between expression divergence and Ks in rice. (C) Fitted smooth spline curves between expression divergence and Ka in Arabidopsis. (D) Fitted smooth spline curves between expression divergence and Ka in rice.
doi:10.1371/journal.pone.0028150.g008

modes (Figure 5 and 8). For WGD duplicates, significant correlations only exist in those generated by recent WGD events - if only relatively 'young' WGD duplicates are studied, the correlations may be overestimated. Moreover, such correlations are not uniformly distributed among Ks levels - at low Ks levels (<1), all duplication modes may show correlations.

We find evidence for duplicate gene retention by both neo-functionalization and genetic redundancy, seemingly at opposite ends of the spectrum of possible fates of duplicated gene pairs. Genetic redundancy has clear biological significance, i.e. provision of buffering capacity [10,11] and/or dosage balance [34,46,47,48], and seems most closely related to WGD or tandem duplicates. The origins of genetic novelty, of clear biological significance in occupation of new niches or adaptation to new environments, may lie more with the greater expression divergence and more independent evolution of distantly transposed and dispersed

**Table 7.** Proportion of pairs of duplicates that have changed DNA methylation status in promoter regions.

| Species | WGD | Tandem duplication | Proximal duplication | DNA based transposed duplication | Retrotransposed duplication | Dispersed duplication |
|---|---|---|---|---|---|---|
| Arabidopsis | 0.303 | 0.290 | 0.309 | 0.387 | 0.347 | 0.318 |
| Rice | 0.357 | 0.417 | 0.404 | 0.416 | 0.447 | 0.385 |

doi:10.1371/journal.pone.0028150.t007

duplications. Proximal duplication is more balanced in its contributions to genetic novelty and redundancy than other gene duplication modes.

Detailed delineation of gene duplication modes reveals some new trends. Prior studies classified genes into as few as two types (anchors generated by polyploidy, and non-anchors generated by single gene duplication [58]), or as many as three types (segmental, tandem and dispersed: [38]). In this study, we have attempted to distinguish DNA/RNA based transposed from dispersed duplication, and proximal from tandem duplication. DNA based transposed duplications tend to evolve faster in expression while having smaller Ks than dispersed duplicates. Tandem duplicates diverge slower in gene expression than proximal duplicates. Proximal duplicates tend to diverge faster in expression than WGD duplicates, though concerted evolution [20] may homogenize their coding sequences.

## The factors that affect expression divergence are complex

Our analyses suggest that it may be inappropriate to make generalizations about levels and patterns of expression divergence across gene duplication modes. Ks, putatively a proxy for age, seems to be related to expression divergence only within a subset of duplication modes and largely only among younger duplicates. Ka, putatively a proxy for functional change, also shows statistically significant and heterogeneous relationships to expression divergence. The level of these correlations is very low, even in recent WGD duplicates.

Although expression divergence between duplicates is often significantly correlated with coding sequence divergence, it is well known that gene expression is also regulated by other genomic regions such as promoters, 5′UTRs, and 3′UTRs. The correlations between expression divergence and nucleotide substitution rates ($\mu$) of different genomic regions for pairs of duplicates are summarized in Table S3. WGD duplicates show significant correlations between expression divergence and nucleotide substitution rates in all three regions. These correlations become marginal and often non-significant among tandem duplicates. Expression divergence of proximal duplicates is more closely associated with divergence in promoters, 5′UTRs and 3′UTRs than coding sequences. Expression divergence of DNA based transposed duplicates seem to be most related to Ka and $\mu$ of 3′UTRs. Expression divergence of dispersed duplicates is very slightly correlated with Ka but not with other substitution rates. Retrotransposed duplication is least related to any type of sequence divergence, consistent with its general separation of a gene from its native regulatory elements.

In partial summary, expression divergence between duplicate genes may be affected by different and multiple genetic factors depending on the causal duplication mechanism. For pairs of orthologs between Arabidopsis and rice, expression divergence seems only correlated with Ka (Table 5 and Table S3). Single gene duplications including translocated and tandem/proximal duplications have been suggested to be much more prone to promoter disruption than WGD [58]. We examined this hypothesis using >45% sequence identity as criterion for determining duplicated (non-disrupted) promoter regions, finding proximal duplicates to have higher proportions of duplicated promoter regions than WGD duplicates (Table 9). This finding seems to contradict the greater expression divergence of proximal duplicates than WGD duplicates. Thus, we note that each of the investigated genetic/epi-genetic factors may only explain a small portion of the variation of expression divergence between duplicate genes, and perhaps only for certain duplication modes. New factors that may affect expression divergence and how different factors work together are worth investigation.

**Table 8.** Proportion of genes that are methylated in promoter regions.

| Species | Singletons | Duplicate genes |
|---|---|---|
| Arabidopsis | 0.185 | 0.157 |
| Rice | 0.224 | 0.217 |

doi:10.1371/journal.pone.0028150.t008

## Possible non-random associations between duplication mode and population size

WGD is often associated with speciation in plants [81,82]. If ancestral polyploidy was attendant with speciation, new species would have likely initially faced very small $N_e$ (i.e. effective population size), weak selection, high drift and high mutational load. This could put a premium on buffering, but allow little chance for beneficial mutations. On the other hand, small-scale duplications may have been only infrequently associated with speciation, if at all. Thus they might be more likely to arise in established populations with larger $N_e$ and more efficient selection, all putting a greater premium on evolutionary novelty to attain fixation. A hypothesis worthy of further investigation is that non-random associations between duplication mode and population size have shaped which specific genes and functional variations are retained.

# Methods

## Genome annotation

Genome annotations were obtained from TAIR (http://www.arabidopsis.org) for Arabidopsis, and from the Rice Genome Annotation Project data (http://rice.plantbiology.msu.edu) for rice. Gene structures were retrieved using ENSEMBL Biomart (http://plants.ensembl.org/biomart/martview).

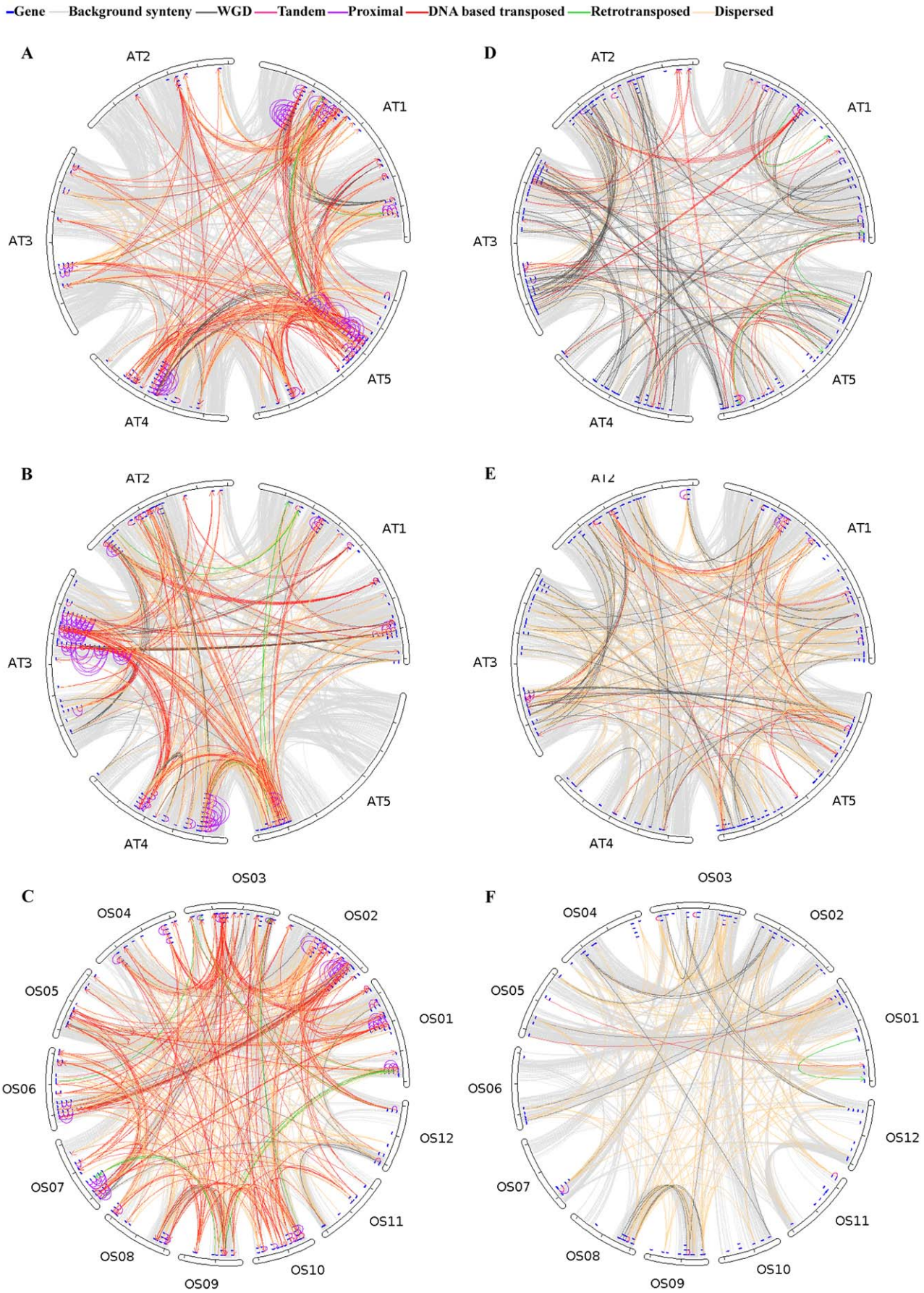
Gene
Background synteny
WGD
Tandem
Proximal
DNA based transposed
Retrotransposed
Dispersed
A
AT2
AT1
AT3
AT5
AT4
D
AT2
AT1
AT3
AT5
AT4
B
AT2
AT1
AT3
AT5
AT4
E
AT1
AT3
AT5
AT4
OS03
C
OS04
OS02
OS05
OS01
OS06
OS12
OS07
OS11
OS08
OS10
OS09
F
OS04
OS02
OS05
OS01
OS06
OS12
OS07
OS11
OS08
OS10
OS09

**Figure 9. Gene duplication modes among the members of selected gene families.** (A) Arabidopsis disease resistance gene homologs. (B) Arabidopsis Cytochrome P450 gene family. (C) Rice Cytochrome P450 gene family. (D) Arabidopsis cytoplasmic ribosomal gene family. (E) Arabidopsis C2H2 zinc finger gene family. (F) Rice C2H2 zinc finger gene family. Different gene duplication modes are indicated by different colors.
doi:10.1371/journal.pone.0028150.g009

## Gene expression data

To reliably assess the expression divergence between duplicates or between orthologs, we used as many publicly available microarray datasets as possible, all of which were obtained from NCBI's GEO (http://www.ncbi.nlm.nih.gov/geo/). At the time of retrieval, 6,009 samples existed for the Affymetrix Arabidopsis ATH1 Genome Array (GEO platform GPL198), of which 800 were not available and a total of 5,209 CEL files were downloaded. 550 CEL files for the Affymetrix GeneChip Rice Genome Array (GEO platform GPL2020) were downloaded, of which 13 were removed due to incorrect array types. For both Arabidopsis and rice raw expression data, RMA normalization was performed using the RMAExpress software (http://rmaexpress.bmbolstad.com) across the entire dataset. Outliers were detected using the arrayQualityMetrics [83] Bioconductor package, which implements three different statistical tests to identify outliers. A total of 443 and 29 samples were detected as outliers and removed in Arabidopsis and rice respectively. Thus, 4,566 and 508 samples remained for Arabidopsis and rice, respectively. The annotation files (Release 30) of these two arrays were downloaded from the Affymetrix website (http://www.affymetrix.com), containing 22,810 Arabidopsis genes and 27,910 rice genes. For a gene, there may be multiple probe sets or multiple types of probe sets available on the array. However, a general rule for selection of a probe set that best represents the gene's expression profile has not been resolved yet [84,85]. In this study, inclusion or exclusion of "sub-optimal" probe sets with suffix "_s_at" or "_x_at" that are suspected of potential cross-hybridization (may be not sub-optimal in practice according to ref. [84,85]) had only trivial effects. Thus, to survey as many genes as possible, all types of probe sets were considered, and for a gene with multiple probe sets, we used the first probe set according to alphabetic sorting to represent its expression profile.

## Analysis of expression data

Similarity between the expression profiles of two duplicate genes within species was initially measured by either Pearson's (denoted by PCC or $r$) or Spearman's correlation coefficient. Note that all replicate chips were retained and correlations were computed across all individual chips. These two measures generated highly consistent results, and thus we only showed the statistics measured by Pearson's correlation coefficient. The expression divergence between two duplicate genes or orthologs was measured by $1-r$ [61,62].

Orthologous gene pairs compared between Arabidopsis and rice were restricted to 2,012 pairs of orthologs located at corresponding loci in paired syntenic blocks between Arabidopsis and rice as identified by MCScan [53], and having expression profiles on the arrays. To assess the expression conservation (EC) for a pair of Arabidopsis-rice orthologs, we adopted a conceptual framework of comparing co-expression patterns across species [69] implemented in several other studies similar to ours [86,87,88,89,90]. In this study, the framework can be described as:

1) The expression matrices, $\mathbf{A}$ and $\mathbf{B}$, in Arabidopsis and rice respectively, are restricted to genes for which orthology relationships have been identified and ordered accordingly (i.e., equivalent rows of the two matrices correspond to the expression profiles of a pair of orthologs):

$$\mathbf{A}=[\mathbf{a}_i]_{i=1,\ldots,k}$$
$$\mathbf{B}=[\mathbf{b}_i]_{i=1,\ldots,k}$$

where $\mathbf{a}_i$ and $\mathbf{b}_i$ are the vectors of expression profiles for any pair $i$ of orthologs for Arabidopsis and rice, respectively, and $k$ is the number of orthologous gene pairs.

2) $\mathbf{A}$ and $\mathbf{B}$ are then converted into two pair-wise correlation matrices, $\mathbf{R}^A$ and $\mathbf{R}^B$, by computing the PCCs between the expression profile of each gene and that of any other gene in each species separately:

$$\mathbf{R}^A=[PCC(\mathbf{a}_i,\mathbf{a}_g)]_{i=1,\ldots,k;g=1,\ldots,k}$$
$$\mathbf{R}^B=[PCC(\mathbf{b}_i,\mathbf{b}_g)]_{i=1,\ldots,k;g=1,\ldots,k}$$

3) The expression conservation for an orthologous gene pair $i$ is computed as:

$$EC(i)=PCC(R^A_{i,g},R^B_{i,g}),\ g=1,\ldots,k$$

Its corresponding expression divergence is $1-EC(i)$.

## Identification of different modes of gene duplications

The populations of potential gene duplications in Arabidopsis or rice were identified using BLASTP. Only the top five non-self protein matches that met a threshold of $E<10^{-10}$ were considered. Genes without BLASTP hits that met a threshold of $E<10^{-10}$ were deemed singletons. Pairs of WGD duplicates were downloaded from the PGDD database [51,53]. Pairs of $\alpha$, $\beta$, $\gamma$ duplicates in Arabidopsis and pairs of $\rho$, $\sigma$ duplicates in rice were obtained from published lists [49,54]. Single gene

**Table 9.** Proportion of copied promoter regions among duplicates.

| Species | WGD | Tandem duplication | Proximal duplication | DNA based transposed duplication | Retrotransposed duplication | Dispersed duplication |
|---|---|---|---|---|---|---|
| Arabidopsis | 0.899 | 0.923 | 0.927 | 0.885 | 0.865 | 0.871 |
| Rice | 0.382 | 0.431 | 0.407 | 0.344 | 0.327 | 0.330 |

doi:10.1371/journal.pone.0028150.t009

duplications were derived by excluding pairs of WGD duplicates from the population of gene duplications. Tandem duplications were defined as being adjacent to each other on the same chromosome. Proximal duplications were defined as non-tandem genes within 20 annotated genes of each other on the same chromosome [38].

The remaining single gene duplications (after deducting tandem and proximal duplications) were searched for distant single gene-transposed duplications. To accomplish this aim, genes at ancestral chromosomal positions need to be discerned by aligning syntenic blocks within and between species [53,55]. Angiosperm syntenic blocks were downloaded from the Plant Genome Duplication Database (PGDD), available at http://chibba.agtec.uga.edu/duplication. At the time of retrieval, PGDD provided syntenic blocks within and between 10 species including *Arabidopsis thaliana*, *Carica papaya*, *Prunus persica*, *Populus trichocarpa*, *Medicago truncatula*, *Glycine max*, *Vitis vinifera*, *Brachypodium distachyon*, *Oryza sativa*, *Sorghum bicolor*, *Zea mays* [51,53]. An Arabidopsis or rice gene locus was regarded as ancestral if the resident gene along with any of its homologous genes (paralogs/orthologs) occur at corresponding loci within any pair of syntenic blocks in PGDD. Using this criterion, the population of Arabidopsis/rice genes was divided into two subsets: genes at ancestral loci and genes that were transposed. For a pair of distantly transposed duplicate genes, we required that one copy was at its ancestral locus and the other was at a non-ancestral locus, named the parental copy and transposed copy respectively. If the parental copy has more than two exons and the transposed copy is intronless, we inferred that this pair of duplicate genes occurred by retrotransposition (RNA based transposition). If both copies have a single exon, the pair of duplicates was unclassified. For other cases of a pair of distantly transposed duplicate genes, we inferred that the duplication occurred by DNA based transposition. The remaining single gene duplications in the population, i.e. after deducting WGD, tandem, proximal, DNA based transposed and retrotransposed duplications from the BLASTP output, were classified as dispersed duplications. After pairs of duplicate genes in each duplication mode were identified, we assigned a unique origin to each duplicated gene, according to the following order of priority: WGD>tandem>proximal>retrotransposed>DNA based transposed>dispersed.

### GO/Pfam enrichment analysis

GO/Pfam enrichment analysis was performed using Fisher's exact test. The *P*-value was calculated for the null hypothesis that there is no association between a subset of genes and a particular functional/domain category and was corrected with the total number of terms to account for multiple comparisons.

### Assessing DNA sequence divergence

Coding sequence divergence between a pair of genes was denoted by either non-synonymous (Ka) or synonymous (Ks) substitution rates. Protein sequences were aligned using Clustalw [91] with default parameters. The protein alignment was then converted to DNA alignment using the "Bio::Align::Utilities" module of the BioPerl package (http://www.bioperl.org/). Ka and Ks were estimated by Nei-Gojobori statistics [92], available through the "Bio::Align::DNAStatistics" module of the BioPerl package. Note that the "Bio::Align::DNAStatistics" module may generate invalid Ka or Ks for some duplicate gene pairs due to mis-alignments, which were ruled out from related analysis. All levels of valid Ka or Ks values were considered in related statistical analyses. Because distributions of Ka or Ks were centered at low levels (~1.0), in related figures, to improve their clarity, we only displayed Ka or Ks values between 0 and 2.0.

The promoter region of a gene was restricted to a maximum of 1,000 bp upstream of the transcription start site (TSS) or less if the nearest adjacent upstream gene is closer than 1,000 bp. For a pair of genes, the divergence of promoter sequences was indicated by their Jukes-Cantor nucleotide substitution rate ($\mu$) [93], which is available through the "Bio::Align::DNAStatistics" module of the BioPerl package. The divergence in 5′UTR and 3′UTR is also measured by nucleotide substitution rates ($\mu$). Note that the "Bio::Align::DNAStatistics" module may not output $\mu$ if the distance between two input nucleotide sequences is too near or too far. Duplicate gene pairs lacking estimation of $\mu$ in the promoter region, 5′UTR or 3′UTR were removed from related analysis.

### DNA methylation data and its analysis

Arabidopsis and rice genome-wide DNA methylation data were obtained from GEO (accession number: GSE21152) [73]. We chose this study, which provided DNA methylation for both Arabidopsis and rice, because the systematic errors between species should be smaller than in data from separate studies. A gene methylated in the promoter region is defined by the presence of two or more adjacent methylated probes within the promoter DNA sequence [59,72].

### Gene families

Lists of published gene families were obtained from TAIR (http://www.arabidopsis.org/browse/genefamily/index.jsp) for Arabidopsis, and from the Rice Genome Annotation Project data (http://rice.plantbiology.msu.edu/annotation_community_families.shtml) for rice. Only families with more than nine genes were considered. Arabidopsis disease resistance gene homologs were downloaded from the NIBLRRS Project website (http://niblrrs.ucdavis.edu/). The Rice Cytochrome P450 gene family was downloaded from the Cytochrome P450 homepage [94] .

## Supporting Information

**Table S1 Enriched GO terms and Pfam domains associated with the duplicates of conserved or divergent expression at each WGD event.**
(DOCX)

**Table S2 List of investigated gene families and their enrichments with modes of gene duplication.**
(DOCX)

**Table S3 Correlations between expression divergence and different types of sequence divergence.**
(DOCX)

## Acknowledgments

We thank Prof. Michael R. Freeling for helpful advice and Barry Marler for IT support. We thank Xinyu Liu for statistical advice.

## Author Contributions

Conceived and designed the experiments: YW XW AHP. Performed the experiments: YW SPF. Analyzed the data: YW XW FAF AHP. Contributed reagents/materials/analysis tools: HT XT. Wrote the paper: YW XW AHP.

## References


1. Paterson AH, Freeling M, Tang H, Wang X (2010) Insights from the comparison of plant genome sequences. Annu Rev Plant Biol 61: 349–372.
2. Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, et al. (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature 431: 946–957.
3. Aury JM, Jaillon O, Duret L, Noel B, Jubin C, et al. (2006) Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature 444: 171–178.
4. Kellis M, Birren BW, Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature 428: 617–624.
5. Wolfe KH, Shields DC (1997) Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387: 708–713.
6. Kassahn KS, Dang VT, Wilkins SJ, Perkins AC, Ragan MA (2009) Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates. Genome Res 19: 1404–1418.
7. Zhang G, Cohn MJ (2008) Genome duplication and the origin of the vertebrate skeleton. Curr Opin Genet Dev 18: 387–393.
8. Ohno S (1970) Evolution by gene duplication. New York: Springer Verlag.
9. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290: 1151–1155.
10. Chapman BA, Bowers JE, Feltus FA, Paterson AH (2006) Buffering of crucial functions by paleologous duplicated genes may contribute cyclicality to angiosperm genome duplication. Proc Natl Acad Sci U S A 103: 2730–2735.
11. VanderSluis B, Bellay J, Musso G, Costanzo M, Papp B, et al. (2010) Genetic interactions reveal the evolutionary trajectories of duplicate genes. Mol Syst Biol 6: 429.
12. Dean EJ, Davis JC, Davis RW, Petrov DA (2008) Pervasive and persistent redundancy among duplicated genes in yeast. PLoS Genet 4: e1000113.
13. DeLuna A, Springer M, Kirschner MW, Kishony R (2010) Need-based up-regulation of protein levels in response to deletion of their duplicate genes. PLoS Biol 8: e1000347.
14. DeLuna A, Vetsigian K, Shoresh N, Hegreness M, Colon-Gonzalez M, et al. (2008) Exposing the fitness contribution of duplicated genes. Nat Genet 40: 676–681.
15. Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, et al. (2003) Role of duplicate genes in genetic robustness against null mutations. Nature 421: 63–66.
16. Kafri R, Dahan O, Levy J, Pilpel Y (2008) Preferential protection of protein interaction network hubs in yeast: evolved functionality of genetic redundancy. Proc Natl Acad Sci U S A 105: 1243–1248.
17. Musso G, Costanzo M, Huangfu MQ, Smith AM, Paw J, et al. (2008) The extensive and condition-dependent nature of epistasis among whole-genome duplicates in yeast. Genome Res 18: 1092–1099.
18. Wang X, Tang H, Bowers JE, Paterson AH (2009) Comparative inference of illegitimate recombination between rice and sorghum duplicated genes produced by polyploidization. Genome Res 19: 1026–1032.
19. Wang X, Tang H, Paterson AH (2011) Seventy million years of concerted evolution of a homoeologous chromosome pair, in parallel, in major Poaceae lineages. Plant Cell 23: 27–37.
20. Wang X, Tang H, Bowers JE, Feltus FA, Paterson AH (2007) Extensive concerted evolution of rice paralogs and the road to regaining independence. Genetics 177: 1753–1763.
21. Song K, Lu P, Tang K, Osborn TC (1995) Rapid genome change in synthetic polyploids of Brassica and its implications for polyploid evolution. Proc Natl Acad Sci U S A 92: 7719–7723.
22. Ozkan H, Levy AA, Feldman M (2001) Allopolyploidy-induced rapid genome evolution in the wheat (*Aegilops-Triticum*) group. Plant Cell 13: 1735–1747.
23. Shaked H, Kashkush K, Ozkan H, Feldman M, Levy AA (2001) Sequence elimination and cytosine methylation are rapid and reproducible responses of the genome to wide hybridization and allopolyploidy in wheat. Plant Cell 13: 1749–1759.
24. Kashkush K, Feldman M, Levy AA (2002) Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. Genetics 160: 1651–1659.
25. Kashkush K, Feldman M, Levy AA (2003) Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. Nat Genet 33: 102–106.
26. O'Neill RJ, O'Neill MJ, Graves JA (1998) Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. Nature 393: 68–72.
27. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, et al. (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature 457: 551–556.
28. Chen ZJ, Pikaard CS (1997) Transcriptional analysis of nucleolar dominance in polyploid plants: biased expression/silencing of progenitor rRNA genes is developmentally regulated in Brassica. Proc Natl Acad Sci U S A 94: 3442–3447.
29. Comai L, Tyagi AP, Winter K, Holmes-Davis R, Reynolds SH, et al. (2000) Phenotypic instability and rapid gene silencing in newly formed arabidopsis allotetraploids. Plant Cell 12: 1551–1568.
30. Lee HS, Chen ZJ (2001) Protein-coding genes are epigenetically regulated in Arabidopsis polyploids. Proc Natl Acad Sci U S A 98: 6753–6758.
31. Rodin SN, Riggs AD (2003) Epigenetic silencing may aid evolution by gene duplication. J Mol Evol 56: 718–729.
32. Rapp RA, Wendel JF (2005) Epigenetics and plant evolution. New Phytol 168: 81–91.
33. Adams KL, Wendel JF (2005) Novel patterns of gene expression in polyploid plants. Trends Genet 21: 539–543.
34. Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, et al. (2005) Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci U S A 102: 5454–5459.
35. Cusack BP, Wolfe KH (2007) Not born equal: increased rate asymmetry in relocated and retrotransposed rodent gene duplicates. Mol Biol Evol 24: 679–686.
36. Freeling M (2009) Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annu Rev Plant Biol 60: 433–453.
37. Zhao XP, Si Y, Hanson RE, Crane CF, Price HJ, et al. (1998) Dispersed repetitive DNA has spread to new genomes since polyploid formation in cotton. Genome Res 8: 479–492.
38. Ganko EW, Meyers BC, Vision TJ (2007) Divergence in expression between duplicated genes in Arabidopsis. Mol Biol Evol 24: 2298–2309.
39. Jiang N, Bao ZR, Zhang XY, Eddy SR, Wessler SR (2004) Pack-MULE transposable elements mediate gene evolution in plants. Nature 431: 569–573.
40. Brunner S, Fengler K, Morgante M, Tingey S, Rafalski A (2005) Evolution of DNA sequence nonhomologies among maize inbreds. Plant Cell 17: 343–360.
41. Kaessmann H, Vinckenbosch N, Long MY (2009) RNA-based gene duplication: mechanistic and evolutionary insights. Nat Rev Genet 10: 19–31.
42. Brosius J (1991) Retroposons - Seeds of Evolution. Science 251: 753–753.
43. Force A, Lynch M, Pickett FB, Amores A, Yan YL, et al. (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151: 1531–1545.
44. Hughes MK, Hughes AL (1993) Evolution of Duplicate Genes in a Tetraploid Animal, *Xenopus-Laevis*. Mol Biol Evol 10: 1360–1369.
45. Hughes AL (1994) The evolution of functionally novel proteins after gene duplication. Proc R Soc Lond B Biol Sci 256: 119–124.
46. Veitia RA (2003) Nonlinear effects in macromolecular assembly and dosage sensitivity. J Theor Biol 220: 19–25.
47. Papp B, Pal C, Hurst LD (2003) Dosage sensitivity and the evolution of gene families in yeast. Nature 424: 194–197.
48. Paterson AH, Freeling M, Tang H, Wang X (2010) Insights from the comparison of plant genome sequences. Annu Rev Plant Biol 61: 349–372.
49. Bowers JE, Chapman BA, Rong JK, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422: 433–438.
50. Paterson AH, Bowers JE, Chapman BA (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proc Natl Acad Sci U S A 101: 9903–9908.
51. Tang H, Bowers JE, Wang X, Ming R, Alam M, et al. (2008) Synteny and collinearity in plant genomes. Science 320: 486–488.
52. Wang X, Shi X, Li Z, Zhu Q, Kong L, et al. (2006) Statistical inference of chromosomal homology based on gene colinearity and applications to Arabidopsis and rice. BMC Bioinformatics 7: 447.
53. Tang H, Wang X, Bowers JE, Ming R, Alam M, et al. (2008) Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. Genome Res 18: 1944–1954.
54. Tang H, Bowers JE, Wang X, Paterson AH (2010) Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. Proc Natl Acad Sci U S A 107: 472–477.
55. Freeling M, Lyons E, Pedersen B, Alam M, Ming R, et al. (2008) Many or most genes in Arabidopsis transposed after the origin of the order Brassicales. Genome Res 18: 1924–1937.
56. Woodhouse MR, Pedersen B, Freeling M (2010) Transposed genes in Arabidopsis are often associated with flanking repeats. PLoS Genet 6: e1000949.
57. Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell 16: 1679–1691.
58. Casneuf T, De Bodt S, Raes J, Maere S, Van de Peer Y (2006) Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant *Arabidopsis thaliana*. Genome Biol 7: R13.
59. Ha M, Kim ED, Chen ZJ (2009) Duplicate genes increase expression diversity in closely related species and allopolyploids. Proc Natl Acad Sci U S A 106: 2295–2300.
60. Li Z, Zhang H, Ge S, Gu X, Gao G, et al. (2009) Expression pattern divergence of duplicated genes in rice. BMC Bioinformatics 10 Suppl 6: S8.
61. Liao BY, Zhang J (2008) Null mutations in human and mouse orthologs frequently result in different phenotypes. Proc Natl Acad Sci U S A 105: 6987–6992.
62. Liao BY, Weng MP, Zhang J (2010) Contrasting genetic paths to morphological and physiological evolution. Proc Natl Acad Sci U S A 107: 7353–7358.
63. Gu ZL, Nicolae D, Lu HHS, Li WH (2002) Rapid divergence in expression between duplicate genes inferred from microarray data. Trends Genet 18: 609–613.

64. Liljegren SJ, Ditta GS, Eshed HY, Savidge B, Bowman JL, et al. (2000) SHATTERPROOF MADS-box genes control seed dispersal in Arabidopsis. Nature 404: 766–770.
65. Briggs GC, Osmont KS, Shindo C, Sibout R, Hardtke CS (2006) Unequal genetic redundancies in Arabidopsis - a neglected phenomenon? Trends Plant Sci 11: 492–498.
66. Vitkup D, Kharchenko P, Wagner A (2006) Influence of metabolic network structure and function on enzyme evolution. Genome Biol 7: R39.
67. Conant GC, Wolfe KH (2007) Increased glycolytic flux as an outcome of whole-genome duplication in yeast. Mol Syst Biol 3: 129.
68. Gu ZL, Rifkin SA, White KP, Li WH (2004) Duplicate genes increase gene expression diversity within and between species. Nat Genet 36: 577–579.
69. Ihmels J, Bergmann S, Berman J, Barkai N (2005) Comparative gene expression analysis by differential clustering approach: application to the *Candida albicans* transcription program. PLoS Genet 1: e39.
70. Chen ZJ, Ni ZF (2006) Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. Bioessays 28: 240–252.
71. Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW, et al. (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. Cell 126: 1189–1201.
72. Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S (2007) Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. Nat Genet 39: 61–69.
73. Feng S, Cokus SJ, Zhang X, Chen PY, Bostick M, et al. (2010) Conservation and divergence of methylation patterning in plants and animals. Proc Natl Acad Sci U S A 107: 8689–8694.
74. Paterson AH, Chapman BA, Kissinger JC, Bowers JE, Feltus FA, et al. (2006) Many gene and domain families have convergent fates following independent whole-genome duplication events in Arabidopsis, Oryza, Saccharomyces and Tetraodon. Trends Genet 22: 597–602.
75. Tan XP, Meyers BC, Kozik A, Al West M, Morgante M, et al. (2007) Global expression analysis of nucleotide binding site-leucine rich repeat-encoding and related genes in Arabidopsis. BMC Plant Biol 7: 56.
76. Mizutani M, Ward E, Ohta D (1998) Cytochrome P450 superfamily in *Arabidopsis thaliana*: isolation of cDNAs, differential expression, and RFLP mapping of multiple cytochromes P450. Plant Mol Biol 37: 39–52.
77. Xu WY, Bak S, Decker A, Paquette SM, Feyereisen R, et al. (2001) Microarray-based analysis of gene expression in very large gene families: the cytochrome P450 gene superfamily of *Arabidopsis thaliana*. Gene 272: 61–74.
78. Jen CH, Manfield IW, Michalopoulos I, Pinney JW, Willats WG, et al. (2006) The Arabidopsis co-expression tool (ACT): a WWW-based tool and database for microarray-based gene expression analysis. Plant J 46: 336–348.
79. Englbrecht CC, Schoof H, Bohm S (2004) Conservation, diversification and expansion of C2H2 zinc finger proteins in the *Arabidopsis thaliana* genome. BMC Genomics 5: 39.
80. Hedges SB, Dudley J, Kumar S (2006) TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics 22: 2971–2972.
81. Stebbins GL (1982) Plant speciation. Prog Clin Biol Res 96: 21–39.
82. Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, et al. (2009) The frequency of polyploid speciation in vascular plants. Proc Natl Acad Sci U S A 106: 13875–13879.
83. Kauffmann A, Gentleman R, Huber W (2009) arrayQualityMetrics–a bioconductor package for quality assessment of microarray data. Bioinformatics 25: 415–416.
84. Elbez Y, Farkash-Amar S, Simon I (2006) An analysis of intra array repeats: the good, the bad and the non informative. BMC Genomics 7: 136.
85. Liao BY, Zhang JZ (2006) Evolutionary conservation of expression profiles between human and mouse orthologous genes. Mol Biol Evol 23: 530–540.
86. Tirosh I, Barkai N (2007) Comparative analysis indicates regulatory neofunctionalization of yeast duplicates. Genome Biol 8: R50.
87. Essien K, Hannenhalli S, Stoeckert CJ, Jr. (2008) Computational analysis of constraints on noncoding regions, coding regions and gene expression in relation to Plasmodium phenotypic diversity. PLoS One 3: e3122.
88. Dutilh BE, Huynen MA, Snel B (2006) A global definition of expression context is conserved between orthologs, but does not correlate with sequence conservation. BMC Genomics 7: 10.
89. Wang Y, Robbins KR, Rekaya R (2010) Comparison of computational models for assessing conservation of gene expression across species. PLoS One 5: e13239.
90. Wang Y, Rekaya R (2009) A comprehensive analysis of gene expression evolution between humans and mice. Evol Bioinform Online 5: 81–90.
91. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680.
92. Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3: 418–426.
93. Jukes TH, Cantor CR (1969) Evolution of protein molecules. New York: Academic Press.
94. Nelson DR (2009) The cytochrome p450 homepage. Hum Genomics 4: 59–65.