Prevalent Role of Gene Features in Determining Evolutionary Fates of Whole-Genome Duplication Duplicated Genes in Flowering Plants[1][W][OA]

Wen-kai Jiang[2], Yun-long Liu[2], En-hua Xia, and Li-zhi Gao*

Key Laboratory of Biodiversity and Biogeography (W.-k.J.) and Plant Germplasm and Genomics Center, Germplasm Bank of Wild Species in Southwest China (W.-k.J., Y.-l.L., E.-h.X., L.-z.G.), Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201, China; and University of the Chinese Academy of Sciences, Beijing 100039, China (Y.-l.L., E.-h.X.)

The evolution of genes and genomes after polyploidization has been the subject of extensive studies in evolutionary biology and plant sciences. While a significant number of duplicated genes are rapidly removed during a process called fractionation, which operates after the whole-genome duplication (WGD), another considerable number of genes are retained preferentially, leading to the phenomenon of biased gene retention. However, the evolutionary mechanisms underlying gene retention after WGD remain largely unknown. Through genome-wide analyses of sequence and functional data, we comprehensively investigated the relationships between gene features and the retention probability of duplicated genes after WGDs in six plant genomes, Arabidopsis (*Arabidopsis thaliana*), poplar (*Populus trichocarpa*), soybean (*Glycine max*), rice (*Oryza sativa*), sorghum (*Sorghum bicolor*), and maize (*Zea mays*). The results showed that multiple gene features were correlated with the probability of gene retention. Using a logistic regression model based on principal component analysis, we resolved evolutionary rate, structural complexity, and GC3 content as the three major contributors to gene retention. Cluster analysis of these features further classified retained genes into three distinct groups in terms of gene features and evolutionary behaviors. Type I genes are more prone to be selected by dosage balance; type II genes are possibly subject to subfunctionalization; and type III genes may serve as potential targets for neofunctionalization. This study highlights that gene features are able to act jointly as primary forces when determining the retention and evolution of WGD-derived duplicated genes in flowering plants. These findings thus may help to provide a resolution to the debate on different evolutionary models of gene fates after WGDs.

Polyploidy, also known as the whole-genome duplication (WGD), plays a significant role in plant diversification and evolution (Otto and Whitton, 2000; Soltis et al., 2009). Recent genome-wide studies have confirmed a number of polyploidy events along different lineages of flowering plants. For instance, it was confirmed that Arabidopsis (*Arabidopsis thaliana*) underwent at least three rounds of WGD (Blanc et al., 2000, 2003; Paterson et al., 2000; Vision et al., 2000; Simillion et al., 2002; Bowers et al., 2003; Maere et al., 2005) and poplar (*Populus trichocarpa*) experienced at least two rounds of WGDs (Tuskan et al., 2006). Polyploidy events are also observed in the soybean (*Glycine max*) genome, which had at least three rounds of WGDs, with the most recent one occurring 13 million years ago (Schmutz et al., 2010). It was shown that the completely sequenced grass species, including rice (*Oryza sativa*), sorghum (*Sorghum bicolor*), *Brachypodium distachyon*, and maize (*Zea mays*), shared at least two rounds of ancient WGDs earlier than 70 million years ago, before the divergence of modern grass lineages (Paterson et al., 2004, 2009; Tang et al., 2010). In addition to these two ancient WGD events, maize underwent one recent round of WGD at approximately 5 to 12 million years ago (Swigonová et al., 2004; Schnable et al., 2009). A recent study showed that all angiosperm species might share two ancient WGD events, one in the common ancestor of extant seed plants and the other in the common ancestor of extant angiosperms (Jiao et al., 2011).

Most newly formed polyploids go through a course of fractionation, including genome rearrangements, gene losses, and epigenetic changes (Sémon and Wolfe, 2007; Doyle et al., 2008). This process starts quickly after the formation of polyploidy and lasts for a long period of evolutionary time, during which duplicated genes have experienced quite distinct evolutionary fates. Theoretical studies insisted that the most likely fate of

[1] This work was supported by the Natural Science Foundation of Yunnan Province (Key Project grant no. 2010CC011), the Top Talents Program of Yunnan Province (grant no. 20080A009), the Hundreds Oversea Talents Program of Yunnan Province, the Hundreds Talents Program of the Chinese Academy of Sciences, the Chinese Academy of Sciences (grant no. KSCX2–YW–N–029), the Chinese Department of Science and Technology (973 Program grant no. 2007CB815703), and a startup grant of the Kunming Institute of Botany, Chinese Academy of Sciences (to L.G.).

[2] These authors contributed equally to the article.

* Corresponding author; e-mail lgao@mail.kib.ac.cn.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Li-zhi Gao (lgao@mail.kib.ac.cn).

[W] The online version of this article contains Web-only data.

[OA] Open Access articles can be viewed online without a subscription.

www.plantphysiol.org/cgi/doi/10.1104/pp.112.200147

duplicated genes is the loss or pseudogenization of one of the duplicates (Walsh, 1995; Lynch and Conery, 2000, 2003). However, several genome-wide analyses revealed that many duplicate copies might survive long after WGDs (Blanc and Wolfe, 2004; Seoighe and Gehring, 2004; Maere et al., 2005), and the overretention of WGD duplicates was shown to be nonrandom in plant genomes across gene families. Some genes are able to duplicate iteratively, whereas others are consistently restored to singleton status along divergent lineages (Seoighe and Gehring, 2004; Paterson et al., 2006). The genes retained in duplicates are not evenly distributed among different functional categories. For instance, genes encoding transcription factors, protein kinases, and ribosomal proteins were found to be overretained after the most recent round of WGD in Arabidopsis (Blanc and Wolfe, 2004; Seoighe and Gehring, 2004; Maere et al., 2005). The pattern of functional bias in duplicated genes can be detected in other lineages as well, such as poplar (Freeling, 2008), rice (Wu et al., 2008), and maize (Schnable et al., 2009). However, studies on the Compositae (Asteraceae) lineage showed that duplicated genes generated by WGD were significantly enriched for genes associated with structural components or cellular organization, and regulatory and developmental genes such as transcription factors were significantly underrepresented. This pattern was almost consistent in all Compositae species investigated (Barker et al., 2008). Together, these observations suggest that some underlying mechanisms might have contributed to the evolutionary fates of duplicated genes derived from WGD events across divergent taxa, although selection forces might be varied substantially among higher taxonomic categories.

Several models have been proposed to shed light on the evolutionary fates of duplicated genes and to explain the nonrandom retention of duplicated genes (Innan and Kondrashov, 2010). Two of the most important scenarios have been widely acknowledged, known as neofunctionalization (Ohno, 1970; Kimura and King, 1979; Walsh, 1995; Force et al., 1999), whereby one of the duplicate copies obtains a new function; and subfunctionalization (Ohno, 1970; Force et al., 1999; Lynch and Force, 2000), whereby each duplicate copy retains a subset of the original set of functions. A number of new models have been recommended and tested, such as buffering for mutations, in which duplicated genes are more robust to deleterious mutations (Chapman et al., 2006); beneficial selection for higher dosage, in which genes with higher dosage are more prone to be duplicated (Kondrashov et al., 2002); and conserved gene retention, in which genes with low evolution rates are prone to be duplicated (Davis and Petrov, 2004; Brunet et al., 2006). However, because most of these mechanisms were inferred from small-scale duplications (SSDs), their practicability to WGD events is questionable. Indeed, comparisons between WGD and SSD events showed that a pattern of anticorrelation in several functional classes enriched for WGD genes was always depleted for SSD genes (Seoighe and Gehring, 2004; Davis and Petrov, 2005; Maere et al., 2005). Thus, exploring evolutionary models applied to WGD events is still an open question to evolutionary biologists.

The evolutionary properties of WGD duplicated genes have been explored in diverse lineages, such as yeast (*Saccharomyces cerevisiae*), Arabidopsis, poplar, *Xenopus laevis*, *Paramecium tetraurelia*, and *Tetraodon nigroviridis* (Papp et al., 2003; Blanc and Wolfe, 2004; Aury et al., 2006; Brunet et al., 2006; Sémon and Wolfe, 2008; Rodgers-Melnick et al., 2012). In addition to the discovery that WGD duplicated genes are functionally biased (Blanc and Wolfe, 2004; Seoighe and Gehring, 2004; Aury et al., 2006; Freeling, 2008; Wu et al., 2008; Schnable et al., 2009), several gene features were identified to associate with their retention probability, such as gene complexity (He and Zhang, 2005), gene length (Chapman et al., 2006), essentiality (He and Zhang, 2006), expression level (Seoighe and Wolfe, 1999), evolutionary rates (Chapman et al., 2006), number of protein interactions (Guan et al., 2007; Hakes et al., 2007), functional category (Blanc and Wolfe, 2004), alternative splicing status (Kopelman et al., 2005), protein structure (Papp et al., 2003; Liang et al., 2008), position in the protein-protein interaction network (Li et al., 2006; Wu and Qi, 2010), and number of phosphorylation sites (Amoutzias et al., 2010). However, it is unlikely that merely one or two features alone can account for the evolutionary process of duplicated genes, as many of these features are correlated to each other. A complex relationship, for example, was found between gene essentiality, protein connectivity, and gene duplicability in yeast and mammals (Prachumwat and Li, 2006; Liang and Li, 2007). Genes with high levels of expression tend to evolve slowly (for review, see Pál et al., 2006). Given the difficulties in determining the relative contributions of these gene features to the retention of duplicated genes after WGD events, integrated analyses of many gene features together in a combined framework may help to disentangle the complicated evolutionary phenomenon of duplicated genes.

Besides the investigation of gene features related to the evolution of WGD duplicated genes, great efforts have been made to elucidate which evolutionary models may apply to WGD duplicates, such as subfunctionalization, neofunctionalization, dosage balance, and beneficial selection for higher dosage (Seoighe and Wolfe, 1999; Papp et al., 2003; Sémon and Wolfe, 2008; Kassahn et al., 2009). Sémon and Wolfe (2008) suggested that subfunctionalization might play a prevalent role in both the initial preservation and long-term evolution of WGD duplicated genes. Another recent study with several vertebrate genomes supported that neofunctionalization is more prevalent (Kassahn et al., 2009). However, Freeling (2008, 2009) argued that if subfunctionalization or neofunctionalization was dominant after WGDs, then functional classes overretained for WGD duplicates might also be enriched for duplicates derived from SSDs, due to the similar evolutionary forces operating on them. Empirical data revealed that the retention pattern is the reverse of this assumption, as a strong anticorrelation was found between these two types of duplicated genes (Seoighe and Gehring, 2004; Davis and Petrov, 2005; Maere et al., 2005). Intriguingly, the dosage balance hypothesis can address this anticorrelation pattern instead. The dosage balance hypothesis, also known as the balance gene drive (Freeling and Thomas, 2006; Birchler and Veitia, 2007, 2010), proposes that the disturbance to the stoichiometric balance among biological network members is under strong negative selection. SSDs for network members are disfavored because they cause imbalance between network members. But for WGDs, duplicated genes under strong dosage balance selection might be overretained after WGDs, as losing one of their copies might cause dosage imbalance. Thus, an inverse pattern of retention would be expected under the dosage balance hypothesis, which corresponds well with empirical data (Seoighe and Gehring, 2004; Maere et al., 2005). Although these hypotheses have improved our understanding of the evolutionary mechanisms of gene retention after WGDs, the relative applicability of these models still remains unknown. Additionally, the importance of each model and/or mechanism may vary among different taxa under investigation; for instance, subfunctionalization would be rare in species with large population sizes (Lynch and Force, 2000). It may also be the case that the evolutionary forces at work depend on the ages of WGDs; for example, the role of gene conservation might decrease with times after WGDs. Thus, the evolutionary process of WGD-derived duplicated genes can be more comprehensively understood when more genomic sequences and expression data are taken into account with an expanded range of studied organisms.

The completions of plant genomes with high quality and the rapidly extended functional data have offered us unprecedented opportunities to study the evolution of duplicated genes generated by WGDs. To the best of our knowledge, a systematic study has not been attempted to trace the evolutionary process of gene retention by integrating all the available genomic sequences and functional data. Using six genomes representing diverse lineages of the flowering plants Arabidopsis, poplar, soybean, rice, sorghum, and maize, we characterized many gene features together and evaluated their relative significance in the evolution of duplicated genes after WGDs through logistic regression analysis. Three major contributors to the retention probability of duplicated genes were identified: structural complexity, evolutionary rates, and GC3 content. Statistical modeling between gene features and retention probability further classified WGD-derived duplicated genes into the three distinct groups. Our investigation of evolutionary behaviors for different groups of duplicated genes suggests that multiple mechanisms, including gene dosage balance, neofunctionalization, and subfunctionalization, may jointly promote the evolution of WGD retained genes, while no single mechanism was found to be dominant over the others. The integrated analyses of diverse sources of sequence and functional data in this study thus shed new light on how retained genes evolved after WGDs in flowering plants.

RESULTS

Identification of Syntenic Blocks Derived from the Most Recent Round of WGD

Over the past decade, the evolutionary history of WGDs has been extensively characterized in the six plant genomes under investigation (Bowers et al., 2003; Paterson et al., 2004; Tuskan et al., 2006; Schnable et al., 2009; Schmutz et al., 2010). However, different methods were separately applied to describing these WGD events from one species to another. This study aimed to examine the patterns of fractionation after WGDs in general across divergent plant lineages; thus, all six genomes, Arabidopsis, poplar, soybean, rice, sorghum, and maize (Supplemental Table S1), were reanalyzed under a similar framework to minimize variations caused by methodology discrepancy. Following Blanc et al. (2003), we classified duplication block pairs into different age groups according to their divergence times estimated from each pair of sister regions. Each pair of synteny genomic regions generated by one duplication event was usually found to include several duplicated gene pairs. We thus calculated the synonymous substitution rates (DS) between each duplicated gene pairs, and then the distribution of DS for these duplicated gene pairs were presented as box plots (for details, see "Materials and Methods"). Two major age groups of syntenic duplication blocks could be clearly identified in Arabidopsis, poplar, soybean, and maize. For rice and sorghum, the duplicated blocks fell into one major age group. These results appear consistent with former studies (Paterson et al., 2004; Tuskan et al., 2006; Schnable et al., 2009; Schmutz et al., 2010). In Arabidopsis, however, we found only two groups of duplication blocks, which is not in agreement with previous results (Simillion et al., 2002; Bowers et al., 2003; Jiao et al., 2011). This inconsistency may be due to the very stringent parameters to identify pairs of duplicated genes of Arabidopsis in this study (with a BLAST parameter of 1e-20; for details, see "Materials and Methods"). As a result, only the α event (the most recent round of WGD in Arabidopsis, as described by Bowers et al. [2003]) could be clearly identified, while the other, older events were missing.

Duplicated blocks derived from ancient WGDs were easily masked by successive evolutionary events; thus, the reconstruction of ancient WGDs appears difficult and error prone. To distinguish the differences between retained and lost duplicated genes, we focused on the youngest syntenic blocks in each species according to the DS distribution plot, which came from the most recent round of WGD, named R1 WGD (Supplemental Fig. S1; for details, see "Materials and Methods"). In this study, blocks of R1 WGDs could be straightforwardly dissected for all three dicot species and one grass species, maize. However, the other two grass species, rice and sorghum, contained several recent segmentally duplicated blocks in addition to the major age groups. These blocks were also treated as R1 blocks in this study, as they were reported to be heavily characterized by gene conversion

(Gao and Innan, 2004; Paterson et al., 2009). The R1 blocks contained quite a few overlaps, which might be caused by small-scale segmental duplications or small fragments left over from ancient WGDs. We resolved these overlapping blocks by removing both small fragments nested in largely duplicated blocks and small, ancient fragments overlapped by those large and young blocks. After following the above-mentioned steps, we collected a relatively clean data set containing segmentally duplicated blocks only from R1 WGD; their positions and relationships are shown in Figure 1. Segmental duplicated blocks from R1 WGDs covered a large genomic proportion varying from 32% to 83% across the different genomes. If only protein-coding genes were taken into account, the coverage of R1 blocks ranges from 50% to 89% among them (Supplemental Table S2).

Comparisons of Biological Features between Retained and Lost Genes

Comparisons between the retained and lost genes are capable of identifying gene features that may promote the probability of gene retention. Duplicate genes with both copies maintained in syntenic blocks after WGD were considered as retained genes (also defined as "ohnologs"), while lost genes are those without any corresponding duplicate copy in syntenic blocks. Features used in this study include gene length, peptide length, intron total length, intron number, intron average length, with or without introns, protein domain diversity, with or without protein domain annotations, average expression level, expression specificity, DS, nonsynonymous substitution rate (DN), selection constraints, number of transcript isoforms, number of alternative splicing isoforms, with or without alternative splicing isoforms, GC content, and GC3 content; the description of each feature can be found in "Materials and Methods." Most features apparently displayed significant differences between the retained and lost genes (Fig. 2). In comparison with the lost genes, retained genes tended to encode longer gene sequences (described as GLEN in Fig. 2 and Supplemental Table S3; $P < 0.01$ in all tests), longer protein-coding regions (described as PLEN in Fig. 2 and Supplemental Table S3; $P < 0.01$ in all tests), and longer introns and more protein

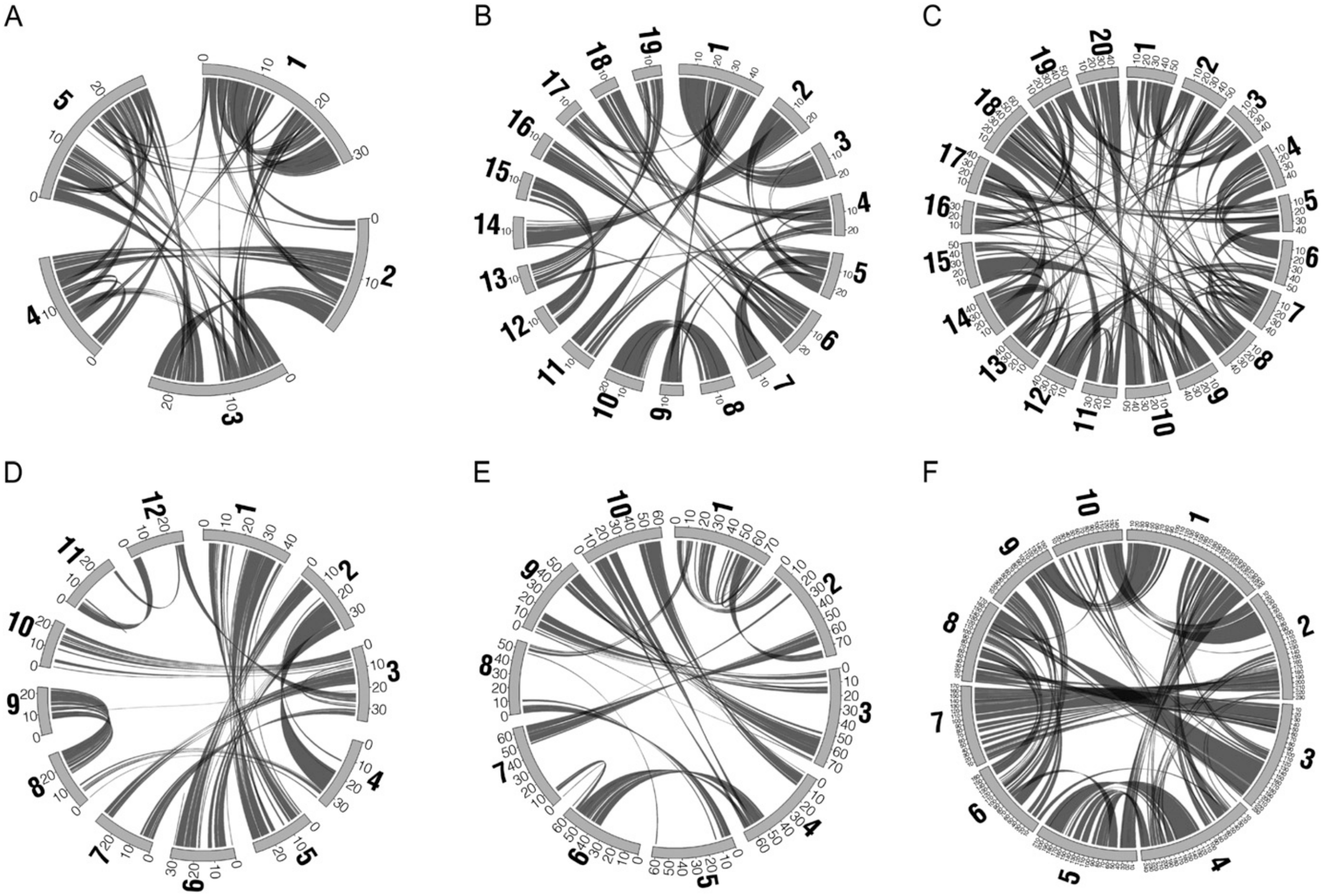

Figure 1. Segmental duplication blocks from the most recent round of WGDs in the six plant genomes. All chromosomes are depicted in a circle, with bands connecting homeolog blocks derived from the most recent round of WGDs in each species. A, Arabidopsis. B, Poplar. C, Soybean. D, Rice. E, Sorghum. F, Maize.

domains (described as INTL and DOMC in Fig. 2 and Supplemental Table S3; $P < 0.01$ in all tests). There were higher fractions of retained genes that contained introns and protein domains, while many lost genes did not contain introns or protein domains at all (described as IN.STATUS and DOM.STATUS, respectively, in Fig. 2 and Supplemental Table S3; $P < 0.01$ in all tests). Thus, retained genes were more complex than lost genes in terms of gene structure. The detection of the regulatory complexity of these retained genes further showed that they were more prone to be alternatively spliced (described as AS and AS.STATUS, respectively, in Fig. 2 and Supplemental Table S3; $P < 0.01$ in all tests), indicative of the complicated nature of their regulatory mechanisms. Apart from gene complexity, retained genes were found to be distinguished from lost genes based on many more biological features. Retained genes were under stronger selection pressures at the sequence level (described as DN/DS in Fig. 2 and Supplemental Table S3; $P < 0.01$ in all tests) and were expressed in more tissues than lost genes (described as EXP.SPE in Fig. 2 and Supplemental Table S3; $P < 0.01$ in all tests except Arabidopsis). In addition, they came with higher expression intensities (described as AVG.EXP in Fig. 2 and Supplemental Table S3; $P < 0.01$ in all tests) and GC3 content (described as GC3 in Fig. 2 and Supplemental Table S3; $P < 0.01$ in all tests).

Different from the above-mentioned patterns of gene features, which were almost consistent across all studied species, evolutionary rates (DN and DS) did not show consistent patterns across the six studied plant genomes (Fig. 2; Supplemental Table S3). Retained genes came with lowered DS in the studied dicot species, whereas elevated DS was detected in all grass species. Moreover, retained genes exhibited lowered DN values in the four genomes of Arabidopsis, soybean, poplar, and maize, while no DN differences could be detected between retained and lost genes in rice and sorghum. Besides variation in evolutionary rates, the pattern of GC content was also found to be distinct among the species under investigation. Retained genes had higher GC content compared with lost genes in soybean, maize, and sorghum. In contrast, retained genes in rice were characteristic of low GC content compared with lost genes, and retained genes possessed similar GC content relative to lost genes in Arabidopsis and poplar.

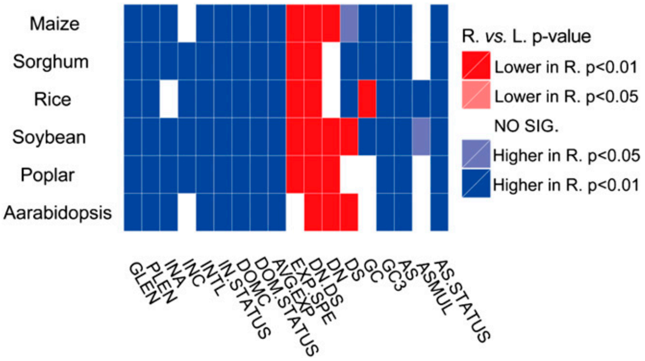

Figure 2. Gene feature comparisons between retained and lost genes. Gene features from the six genomes are compared between WGD retained and lost genes through Wilcox's rank sum test or Fisher's exact test. Each cell shown represents one statistical test (retained versus lost). Blue boxes indicate that the value for that gene feature is significantly higher in retained genes than in lost genes, while red boxes denote lower values. Features are abbreviated as follows: GLEN, gene length; PLEN, peptide length; INA, intron average length; INTL, intron total length; IN.STATUS, with or without introns; DOMC, protein domain number; DOM.STATUS, with or without protein domain annotation; AVG.EXP, average expression value; EXP.SPE, expression specificity; DN, nonsynonymous evolution rate; DS, synonymous evolution rate; DN.DS, DN/DS ratio; GC, GC percentage; GC3, GC percentage at third codon position; AS, number of gene isoforms; ASMUL, number of alternative splicing isoforms (excluding genes without alternative splicing isoforms); AS.STATUS, with or without alternative splicing isoforms. Full statistical test results for each comparison can be found in Supplemental Table S3.

Additional analyses showed that patterns of alternative splicing also varied among these species. Retained genes were more likely to be alternatively spliced (described as AS and AS.STATUS in Fig. 2 and Supplemental Table S3), resulting in more transcript isoforms. However, when merely considering those containing alternative splicing isoforms, retained genes failed to show high levels of splicing complexity in the majority of studied species except rice and soybean (described as ASMUL in Fig. 2 and Supplemental Table S3). Retained genes in these two species were not only more prone to being alternatively spliced (described as AS and AS.STATUS in Fig. 2 and Supplemental Table S3; $P < 0.01$ for both species) but also tended to have more splicing isoforms per gene compared with lost genes (described as ASMUL in Fig. 2 and Supplemental Table S3; $P < 0.01$ for rice, $P < 0.05$ for soybean). For the remaining four studied species (Arabidopsis, poplar, sorghum, and maize), duplicated genes with alternative splicing isoforms were preferably retained (described as AS and AS.STATUS in Fig. 2 and Supplemental Table S3; $P < 0.01$ in all tests). However, the number of isoforms per gene created by alternative splicing did not correspondingly increase the retention probability of duplicated genes (described as ASMUL in Fig. 2 and Supplemental Table S3; $P > 0.05$ for all four species). Thus, it is the ability of being alternatively spliced rather than the diversity of alternative isoforms that might be more relevant to gene retention probability. The observation of positive correlation between AS and the retention probability of duplicated genes did not correspond with previous studies in animals, where a negative correlation was reported (Kopelman et al., 2005; Su et al., 2006; Talavera et al., 2007; Jin et al., 2008). Note that these studies used either all duplicated genes or tandemly duplicated genes in the genome, while our analysis merely included WGD retained genes. To make a comparison, we investigated the relationship between AS and tandem duplicated genes and did find a negative correlation between them (Supplemental Fig. S2). In Arabidopsis, for example, genes with more AS isoforms came with higher proportions of WGD-derived duplicated genes (Supplemental Fig. S2A; χ^2 test, $P < 0.01$). In reverse, genes with more AS isoforms rarely underwent tandem duplications (Supplemental Fig. S2B; χ^2 test, $P < 0.01$). Thus, alternative splicing might play a different role in duplicated genes retained after WGDs versus other sources of gene duplications. Gene features between duplicated genes generated by WGDs and tandem duplicated genes created through SSDs were further compared to investigate whether they exhibited similar behaviors. The results showed that WGD retained genes were structurally more complex, evolved more slowly, were expressed at higher levels, and came with higher GC3 contents in comparison with tandem gene duplications (Supplemental Fig. S3).

Major Contributors to the Probability of Gene Retention

As gene features relevant to the retention probability of duplicated genes were highly correlated with each other (Supplemental Fig. S4), logistic regression modeling was carried out based on principal component analysis (PCA) to reduce the complexity and exclude problems of parameter correlation. PCA grouped several features together that were consistent among the studied genomes (Supplemental Table S4). Here, component 1 was treated as gene structure complexity, with which gene length, total intron length, and number of introns were highly correlated. Component 2 may be regarded as selection pressure, in which DN and DN/DS dominated. Component 3 was mainly affected by average intron length, which may be noted as "gene compactness" with a tightly genic structure. Component 4 can be considered as patterns of gene expression, in which expression intensity and specificity were highly correlated; that is, component 4 was positively correlated to expression level but negatively correlated to expression specificity. We observed that component 2 could not be easily distinguished from component 4. In Arabidopsis, for instance, correlation values between component 2 and DN and DN/DS were 0.826 and 0.836, respectively. However, component 2 was also highly correlated to both expression intensity and specificity, with correlation values of −0.694 and 0.558, respectively. Thus, component 2 was a mixture of selection pressure and expression patterns, in which selection pressures dominated. Similarly, component 4 combined selection pressure and patterns of expression together, in which patterns of expression dominated (Supplemental Table S4). This observation is not surprising, due to the high correlation between expression patterns and evolutionary rates realized by numerous empirical studies (for review, see Pál et al., 2006). To this end, we grouped nine gene features into the four combined parameters, which dramatically simplified our analyses.

To find out the major contributors determining the probability of gene retention, we performed logistic regression analyses based on PCA. The principal components generated by PCA were combined with the remaining features, including the number of gene isoforms, number of domains, DS, as well as GC and GC3 contents. All these features were afterward subjected to logistic regression analyses between retained and lost genes. Although regression results differed among these species, the three contributors of GC3 content, selection pressure (component 2), and domain number were observed to be significant across all studied species (Table I; $P < 0.05$ for all tests). However, it seems that the remaining features, such as component 3 (gene compactness) and component 4 (patterns of gene expression) played limited roles in gene retention, while the isoform diversity of genes turned out to be completely ineffectual.

Classification of Retained Duplicated Genes Based on Their Biological Features

Our results of logistic regression analysis may suggest that the three major types of features are related to the probabilities of gene retention: that is, gene complexity, selection pressure for genes, and gene GC3 content. To confirm this pattern, a k-means clustering for WGD retained genes was performed on the basis of parameters related to these three features: gene length, peptide length, intron length, average intron length, number of introns, protein domain diversity, average

Table I. *Logistic regression based on PCA analysis*

Principal component output by PCA and remaining gene features were subjected to logistic regression (retained versus lost). β denotes the regression coefficient, and *P* denotes the significance for that regression coefficient. Factors with significant contributions are in boldface. Component 1, Gene structural complexity; component 2, selection pressure; component 3, gene compactness; component 4, expression pattern. Factors that are significant after regression are indicated with asterisks: **significant at the 0.01 level; *significant at the 0.05 level.

Parameter	Arabidopsis		Soybean		Poplar		Rice		Sorghum		Maize	
	β	P	β	P	β	P	β	P	β	P	β	P
Component 1	**−0.094**	**0.001****	0.029	0.156	**0.098**	**0****	0.005	0.889	−0.018	0.623	**0.086**	**0****
Component 2	**−0.091**	**0****	**−0.238**	**0****	**−0.169**	**0****	**−0.056**	**0.048***	**−0.075**	**0.002***	**−0.126**	**0****
Component 3	0.024	0.509	**−0.096**	**0.001***	**0.057**	**0.038***	0.057	0.092	0.015	0.688	0.031	0.164
Component 4	**−0.102**	**0.003****	−0.014	0.669	**−0.176**	**0****	0.029	0.474	**0.079**	**0.012***	0.006	0.785
DS	**−0.181**	**0.01****	**−0.369**	**0****	−0.037	0.542	0.118	0.085	**0.108**	**0.039***	−0.04	0.31
No. of gene isoforms	0	1	0.023	0.583	0.063	0.129	0.064	0.059	0.063	0.362	−0.023	0.155
Protein domain number	**0.051**	**0.002****	**0.038**	**0.011***	**0.039**	**0.002****	**0.061**	**0****	**0.069**	**0****	**0.028**	**0.02***
GC	−0.325	0.421	**0.712**	**0.011***	**3.735**	**0****	0.293	0.368	0.408	0.212	0.156	0.403
GC3	**0.467**	**0.016***	**0.382**	**0.002****	**0.321**	**0.044***	**0.417**	**0.003****	**0.344**	**0.019***	**0.334**	**0****

expression level, expression specificity, GC content, GC3 content, DN, DN/DS, and DS. We excluded parameters related to alternative splicing in cluster analysis, as they did not correlate to the probabilities of gene retention. The k-means clustering algorithm requires the calculation of means for variables (for details, see "Materials and Methods") and thus is most suitable for continuous variables. The information for two binary parameters can be covered by other parameters (IN.STATUS by intron and intron number, and DOM.STATUS by protein domain diversity). Genes without the annotation of protein domain, for example, are coded as "0" in the variable "protein domain diversity." To simplify the analysis, we removed two binary parameters (IN.STATUS and DOM.STATUS) from further analysis. It was shown that retained genes were clearly classified into three groups by k-means clustering analysis. This finding could be confirmed by some cluster validation criteria, such as Silhouette width, Dunn index, and average distance between means (data not shown). The k-means clustering results showed that the characteristics of the three groups appeared in good accordance with the three major contributors deduced by regression analysis (Supplemental Tables S5–S10). Using Arabidopsis as an example, we found that type I retained genes came with low evolutionary rates (DN and DN/DS in Supplemental Table S5), high expression level (AVG.EXP in Supplemental Table S5), and very low expression specificity (Supplemental Table S5). Type II was mainly determined by structural complexity; they encode longer gene sequences, longer introns, longer peptide sequences, and more protein domains (GLEN, PLEN, INTL, INC, and DOMC in Supplemental Table S5). Type III was represented by high GC/GC3 content and high levels of DS (GC, GC3, and DS in Supplemental Table S5).

To investigate the characteristics of these three groups of retained genes, besides the above-mentioned 13 features used for clustering analysis, all remaining gene features were also included and compared among them. Since this study aimed to uncover the reasons for the preferential retention of WGD retained genes, tandem duplicated and lost genes were included for comparisons as well. The mean value for each gene feature in all these groups was normalized to a 0 to 1 interval, then a heat map was drawn to investigate the whole data set in a concise manner (Fig. 3). Type I genes were always the most conserved, with the lowest values of DN and DS, and were highly and broadly expressed. Retained genes belonging to type II were the most structurally complex in general. They always had the longest genes, proteins, and introns, possessed the greatest number of introns and protein domains, and contained more gene isoforms. Type III genes came with high GC/GC3 contents, elevated evolutionary rates, lower expression levels, strong tissue-specific expression patterns, and less complex structures. Note that all three types of duplicated genes can be distinguished from both lost and tandemly duplicated genes in terms of one of the three major features of evolutionary rate, structural complexity, and GC/GC3 content. Among them, type I and type II genes could be identified from lost genes straightforwardly, while type III genes seemed to be similar to tandem duplicated genes except that they possessed very high GC/GC3 content.

Evolution of Retained WGD Genes

The classification of retained WGD genes into three distinct groups demonstrates that the retention of these genes may be explained by their unique features compared with lost genes. They could be retained because they were structurally very complex, evolved slowly, or came with high GC3 content. Since these types of retained WGD genes came with dissimilar features, we further traced their evolutionary behaviors to ascertain why these features correlate to the probability of gene retention.

We first estimated sequence divergence between the two copies derived from the same WGD event. Patterns of sequence divergence were distinct for ohnologs (duplicated gene pairs generated by WGD) belonging to different types of WGD retained genes, even though they were created by the same WGD event. Compared with ohnologs from type III, both type I and type II diverged more slowly, as measured by DN or DS (Fig. 4; Supplemental Fig. S5). Among them, members of type I had the smallest nonsynonymous level. Distinct patterns were observed between dicot and grass species in terms of DN/DS comparisons. In dicot species, ohnologs from type I diverged the slowest, while those from type III evolved the quickest (Supplemental Fig. S6, A–C). In grass species, however, ohnologs from both type I and type III diverged slowly after WGDs, while ohnologs from type II diverged most quickly (Supplemental Fig. S6, D–F).

We next examined the extent of positive selection of retained genes after WGD. Interestingly, we observed that type I genes were rarely under positive selection, while type III genes came with strong evidence of positive selection after WGD. In Arabidopsis, for instance, type I genes had 7.28% amino acid sites under positive selection after WGD on average, while type III genes had 11.23% sites under positive selection on average (Wilcox's rank-sum test, $P < 0.01$; Fig. 5A). This finding held true in the other studied plant species (Fig. 5).

We then moved to compare the expression correlation coefficient of ohnologs among the three types of WGD retained genes and found no apparent patterns of expression divergence. Ohnologs from type I showed higher expression correlation coefficients in Arabidopsis and poplar (Supplemental Fig. S7, A and B), indicative of slight divergence at the expression level. However, this was not the case in other studied plants. Type II ohnologs diverged the slowest in grass species (Supplemental Fig. S7, D–F), while no difference could be found in soybean (Supplemental Fig. S7C). Calculating expression correlation coefficients may be misleading due to the high error

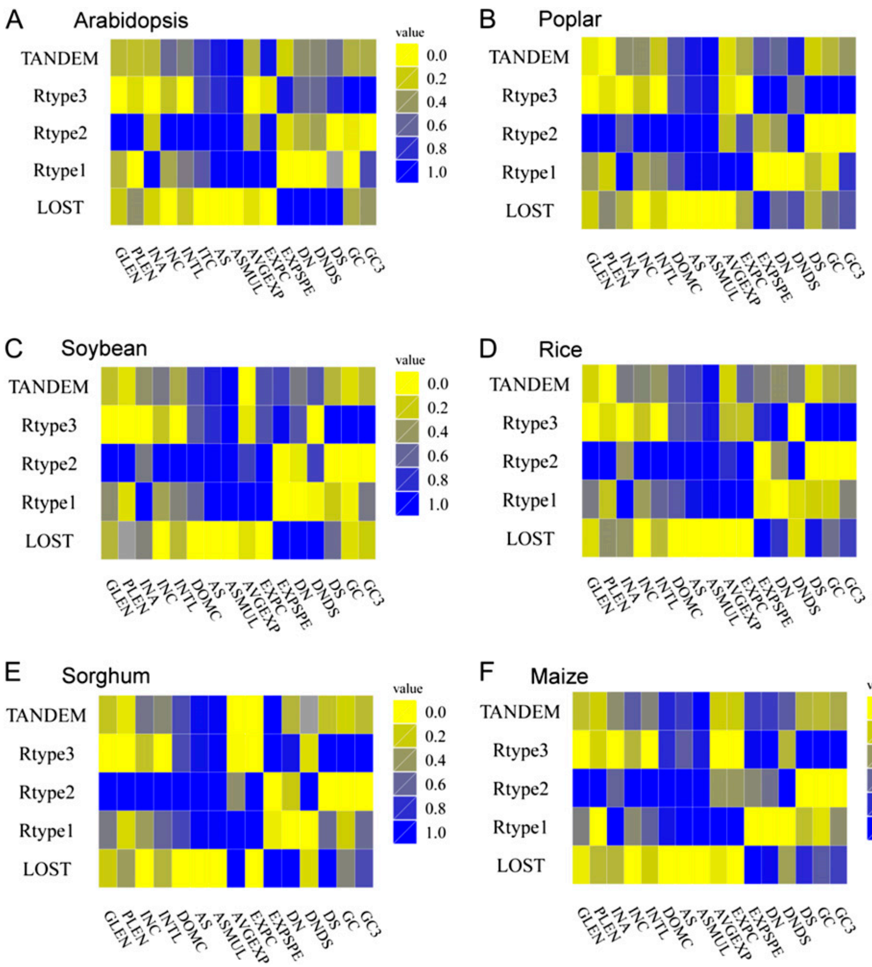

Figure 3. Heat map for gene feature comparisons between types of WGD retained genes. Each cell represents mean values (or percentage) for one gene feature in one type of genes. Values are normalized to a 0 to 1 interval through the following formula: (value − minimum value)/(maximum value − minimum value). Colors are coded by normalized value gradients. TANDEM, Tandem duplicated genes; Rtype1, type 1 WGD retained genes; Rtype2, type 2 WGD retained genes; Rtype3; type 3 WGD retained genes; LOST, genes lost after WGD. Gene feature abbreviations are as follows: GLEN, gene length; PLEN, peptide length; INA, intron average length; INTL, intron total length; AVG.EXP, average expression value; EXPC, number of expressed tissues; EXP.SPE, expression specificity; DN, nonsynonymous evolution rate; DS, synonymous evolution rate; DN.DS, DN/DS; GC, GC percentage; GC3, GC percentage at third codon position; AS, number of gene isoforms; ASMUL, number of alternative splicing isoforms (excluding genes without alternative splicing isoforms). A, Arabidopsis. B, Poplar. C, Soybean. D, Rice. E, Sorghum. F, Maize.

rates of expression intensity measurement methods of microarrays or ESTs. Hence, we also compared spatial expression domains between ohnologs. Expression patterns for each gene were simplified as present or absent across diverse tissues. To discriminate expression divergence among the three types of retained WGD genes, we simply classified patterns of expression divergence into three major types: “similar” indicates that both copies were expressed in most of the tissues; “complement” refers to complementary expression; and “asymmetric” indicates that one copy was expressed in most of the tissues while the other was not expressed or was expressed in very few tissues (for details, see “Materials and Methods”). Such a simplification allowed us to observe patterns of expression divergence between the three types of retained WGD genes. Type I genes contained the highest percentage of ohnologs that had similar expression patterns, while type III genes contained the fewest ohnologs of similar expression. Take Arabidopsis as an example: 73.03% of ohnologs from type I were similarly expressed, but this percentage decreased to only 24.73% for type III (Table II). On the other hand, type III constituted the highest percentage of ohnologs with asymmetric expression, while type I had the lowest (72.71% for type III, 26.97% for type I; data from Arabidopsis in Table II). The pattern held the same in Arabidopsis when we excluded gene pairs with one of their copies that could not be detected by microarray or EST measures (Supplemental Table S11).

We finally examined the relationship between WGD-generated and tandem duplicated genes. It was observed that, in all six studied genomes, both type I and type II genes were rarely duplicated by tandem duplication, while type III genes were highly duplicated through tandem duplication (Table III).

Functional Classes of WGD Retained Genes

The functional roles of different types of WGD retained genes were explored through Gene Ontology (GO) enrichment analysis. Interestingly, these types of genes showed distinct functional roles, and the pattern held true across all studied plant species (Table IV). Type I genes made major contributions to central biological processes, such as translation (GO:0006412), regulation of transcription (GO:0045449), and protein catabolic process (GO:0030163). Type II genes were those related to cell structure construction and cellular metabolic processes, such as microtubule-based movement (GO:0007018), microtubule motor activity (GO:0003777),

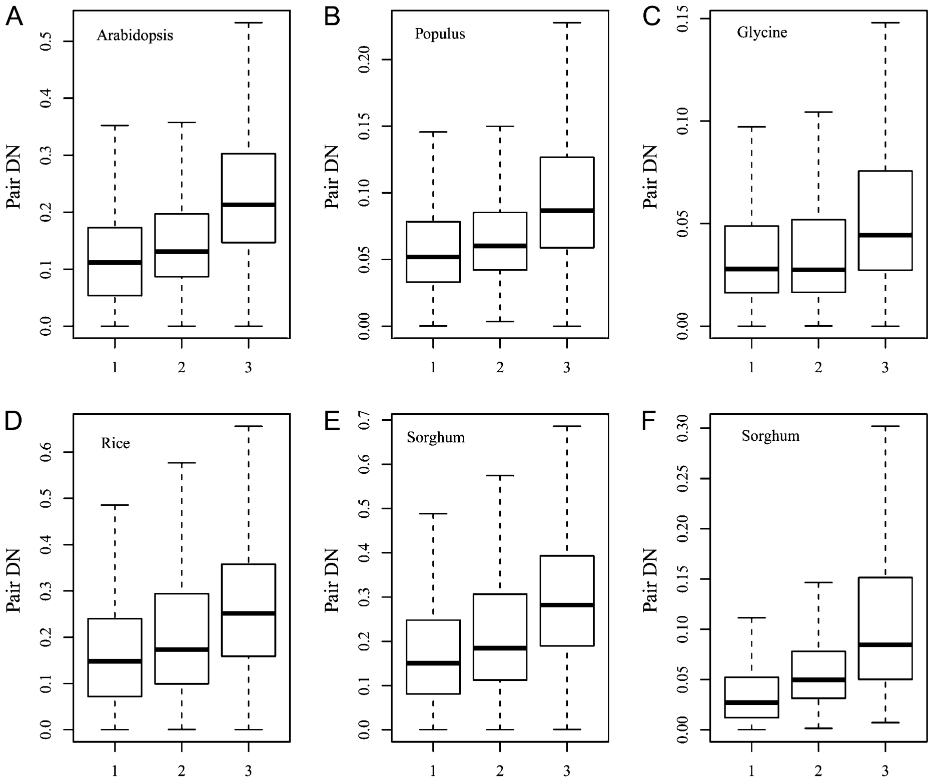

Figure 4. Pairwise DN difference between different types of WGD retained ohnologs. DN was calculated between gene pairs retained after WGD; thus, the DN was "pairwise DN," indicating the divergence rate between the two duplicated copies.

glycolysis (GO:0006096), and cellular amino acid biosynthetic process (GO:0008652). Type III genes had more specific functional roles and were enriched for regulation of transcription, regulation of transcription, DNA dependent, and sequence-specific DNA-binding transcription factor activity (Table IV). Hence, they shared some functional roles with type I genes but carried out more specific functions.

The differences between these three types of WGD retained genes were further investigated in Arabidopsis. There were 66 GO categories enriched for type I genes, 63 enriched for type II genes, and 28 enriched for type III genes (Supplemental Table S12). Only three GO categories (DNA binding; regulation of transcription, DNA dependent; and nucleus) were shared by all three types of retained genes (Supplemental Table

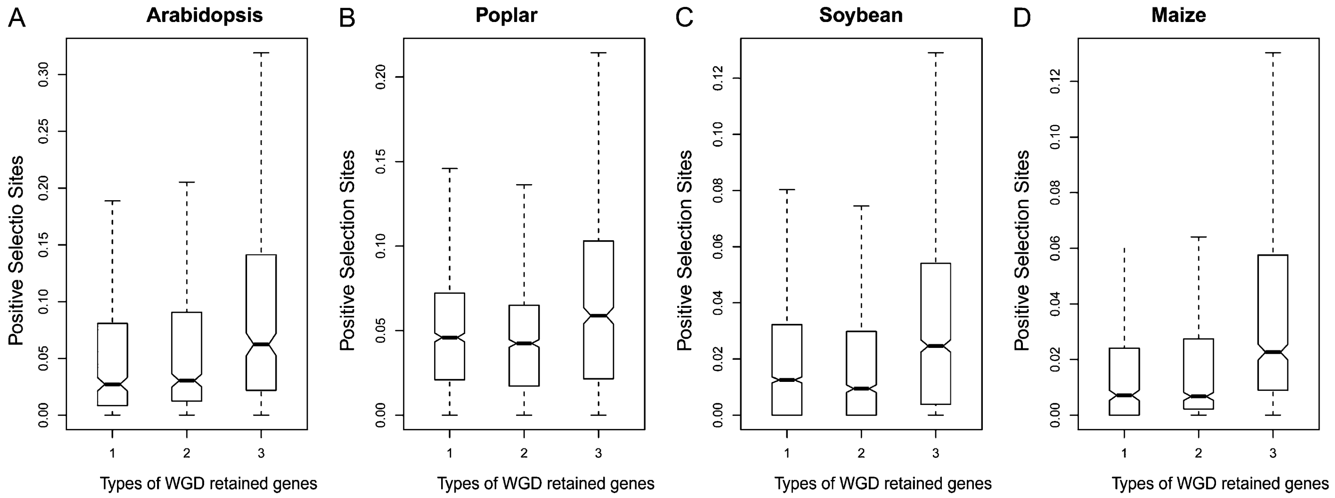

Figure 5. Extent of positive selection for different types of WGD retained genes. The percentage of amino acid sites under positive selection in each gene was plotted against retained gene types. A, Arabidopsis (compared with *A. lyrata*). B, Poplar (compared with cucumber [*Cucumis sativus*]). C, Soybean (compared with *Medicago truncatula*). D, Maize (compared with sorghum).

Table II. *Tissue expression pattern divergence between pairs of duplicates for type I and type III genes*

Expression pattern (expressed or not in each tissue) between duplicated gene pairs is compared between type I and type III WGD retained genes. Similar, Both gene copies express in most of the tissues; asymmetric, one gene expresses in most of the tissues, while the other copy does not express or is expressed in very few tissues.

Expression	Rtype1	Rtype3	χ^2 Test
		%	
Arabidopsis			
Similar	73.03	24.73	$P < 0.001$
Asymmetric	26.97	72.71	
Poplar			
Similar	57.06	29.78	$P < 0.001$
Asymmetric	42.35	64.28	
Soybean			
Similar	94.10	81.80	$P < 0.001$
Asymmetric	5.10	15.10	
Rice			
Similar	49.82	28.79	$P < 0.001$
Asymmetric	49.08	68.52	
Sorghum			
Similar	10.03	4.60	$P < 0.001$
Asymmetric	69.57	81.90	
Maize			
Similar	69	12.63	$P < 0.001$
Asymmetric	21	66.71	

S13). Type I genes generally contained genes that were fundamental to the survival of plants, including genes in central genetic and metabolic networks, such as macromolecular complex (GO:0032991), metabolic process (GO:0008152), and structural constituent of ribosome (GO:0003735). They also comprised genes in response to major environmental changes, such as response to water deprivation (GO:0009414), response to freezing (GO:0050826), and defense response to bacterium (GO:0042742). Additionally, they could participate in several central genetic regulation processes, such as regulation of biological process (GO:0050789), regulation of cellular process (GO:0050794), and transcription factor binding (GO:008134). Compared with type I, type II genes had more specific functional roles, such as regulation of catabolic process (GO:0009894), regulation of cell cycle (GO:0051726), cellulose metabolic process (GO:0009894), glycolysis (GO:0006096), phosphatase activity (GO:0016791), kinase activity (GO:0016301), and helicase activity (GO:0004386). As for type III genes, most of their functional roles were relevant to specific environmental stimuli, such as response to GA stimulus (GO:009739), response to chitin (GO:0010200), response to salicylic acid stimulus (GO:0009751), ethylene-mediated signaling pathway (GO:0009873), protein disulfide oxidoreductase activity (GO:0015035), and pectinesterase activity (GO:0030599).

The dosage balance hypothesis predicts that genes preferred to be retained after WGD are less likely to be duplicated through SSDs, which is supported by the results obtained in this study. Of 157 GO categories enriched for three types of WGD retained genes in Arabidopsis, only 14 were also enriched for tandemly duplicated genes, while up to 56 were indeed depleted for tandem duplicate genes (Fisher's exact test, $P < 0.0001$; Table IV; Supplemental Table S12). To determine whether the pattern held across all six plant genomes under investigation, comparisons were further carried out for all three types of WGD retained genes, showing that they were rarely duplicated through tandem duplications (Table V).

DISCUSSION

The pervasive role of polyploidy in the evolution of eukaryotes has been continuously documented during the last decade (Otto and Whitton, 2000; Wendel, 2000; Jaillon et al., 2004; Kellis et al., 2004; Dehal and Boore, 2005; Cui et al., 2006; Wood et al., 2009). Despite the prevalence of polyploidization in diverse lineages, the mechanisms underlying the evolutionary fates of the retained duplicate genes are still unsolved. Recent studies have characterized several gene features that may be relevant to the probability of gene retention (Seoighe and Wolfe, 1999; Papp et al., 2003; He and Zhang, 2005, 2006; Kopelman et al., 2005; Aury et al., 2006; Chapman et al., 2006; Li et al., 2006; Liang et al., 2008; Wu and Qi, 2010). To date, many features found to be correlated with the probability of gene retention were studied in only one or two species, and their applicability to flowering plants still largely remains to be confirmed. Our analyses expanded the investigation into six genomes of flowering plants by evaluating a total of 18 gene features. Comparisons of gene features between retained and lost genes highlight that a number of them are highly correlated to the probability of gene retention after WGDs. The inconsistency between single-parameter and regression analyses was not unexpected, as many gene features are indeed highly correlated to each other. One feature's effect might be totally encompassed by another, and regression analysis could only identify features that are dominant during

Table III. *Frequencies of tandemly duplicated genes (TAN) in each type of WGD retained gene*

Species	Type I TAN Frequency	Type II TAN Frequency	Type III TAN Frequency	Lost TAN Frequency
Arabidopsis	80/1,919 (4.2%)	92/1,929 (4.8%)	173/1,929 (9.0%)	763/14,409 (5.3%)
Poplar	238/6,073 (3.9%)	161/5,131 (3.1%)	392/5,215 (7.5%)	758/17,205 (4.2%)
Soybean	492/11,775 (4.2%)	300/9,324 (3.2%)	403/8,428 (4.8%)	493/11,447 (4.1%)
Rice	75/1,932 (3.9%)	53/1,494 (3.5%)	167/1,888 (8.8%)	545/14,628 (3.6%)
Sorghum	91/1,383 (6.6%)	32/1,416 (2.3%)	162/1,717 (9.4%)	555/12,924 (4.1%)
Maize	61/1,710 (3.6%)	57/2,493 (2.3%)	188/2,982 (6.3%)	621/17,171 (3.6%)

Table IV. *GO categories enriched for the three types of WGD retained genes in at least four of the six analyzed plant species*

All categories listed are significantly enriched for one type of WGD retained gene by hypergeometric test ($P < 0.01$). BP, Biological process; CP, cellular component; MF, molecular function.

GO Identifier	Type	GO Description	GO Root
GO:0030163	Type I	Protein catabolic process	BP
GO:0045449	Type I	Regulation of transcription	BP
GO:0006412	Type I	Translation	BP
GO:0051246	Type I	Regulation of protein metabolic process	BP
GO:0005840	Type I	Ribosome	CP
GO:0003735	Type I	Structural constituent of ribosome	MF
GO:0030528	Type I	Transcription regulator activity	MF
GO:0043565	Type I	Sequence-specific DNA binding	MF
GO:0003700	Type I	Sequence-specific DNA-binding transcription factor activity	MF
GO:0006096	Type II	Glycolysis	BP
GO:0008652	Type II	Cellular amino acid biosynthetic process	BP
GO:0006468	Type II	Protein amino acid phosphorylation	BP
GO:0007018	Type II	Microtubule-based movement	BP
GO:0016773	Type II	Phosphotransferase activity, alcohol group as acceptor	MF
GO:0005524	Type II	ATP binding	MF
GO:0004674	Type II	Protein Ser/Thr kinase activity	MF
GO:0016301	Type II	Kinase activity	MF
GO:0004713	Type II	Protein Tyr kinase activity	MF
GO:0005515	Type II	Protein binding	MF
GO:0003777	Type II	Microtubule motor activity	MF
GO:0090304	Type III	Nucleic acid metabolic process	BP
GO:0045449	Type III	Regulation of transcription	BP
GO:0006355	Type III	Regulation of transcription, DNA dependent	BP
GO:0005634	Type III	Nucleus	CP
GO:0003677	Type III	DNA binding	MF
GO:0030528	Type III	Transcription regulator activity	MF
GO:0043565	Type III	Sequence-specific DNA binding	MF
GO:0003676	Type III	Nucleic acid binding	MF
GO:0030599	Type III	Pectinesterase activity	MF
GO:0003700	Type III	Sequence-specific DNA-binding transcription factor activity	MF
GO:0015035	Type III	Protein disulfide oxidoreductase activity	MF
GO:0008270	Type III	Zinc ion binding	MF

the evolution process. Besides those proposed by other authors, this study reports some new features correlated to the probability of gene retention, such as intron composition (intron number and intron length) and GC3 content, in all six plant genomes under investigation. However, when they were integrated together through PCA and logistic regression, only three gene features, GC3 content, evolutionary rate, and gene complexity, were found to correlate to the probability of gene retention.

Of the three major contributors inferred by regression analysis, the role of gene conservation in the process of gene retention indeed corresponds to previous observations, in which it was found that genes

Table V. *Enrichment of WGD retained genes versus tandem duplicated genes in GO categories*

Total denotes the number of GO categories enriched for one type of WGD retained genes; TAN Over denotes the number of GO categories enriched for both WGD retained and tandem duplicated genes; TAN Under denotes the number of GO categories enriched for WGD retained genes while depleted for tandem duplicated genes; Over versus Under denotes *P* values of statistical tests for comparisons of frequency distribution between two types of GO categories (TAN Over versus TAN Under).

Species	Type I				Type II				Type III			
	Total	TAN Over	TAN Under	Over versus Under	Total	TAN Over	TAN Under	Over versus Under	Total	TAN Over	TAN Under	Over versus Under
Arabidopsis	66	2/66	23/66	<0.01	63	9/63	23/63	<0.01	28	3/28	10/28	0.055
Soybean	31	3/31	13/31	<0.01	49	5/49	19/49	<0.01	29	3/29	11/26	<0.05
Poplar	48	0/48	25/42	<0.01	57	10/57	20/57	0.0546	23	5/23	9/23	0.337
Rice	20	0/20	13/20	<0.01	29	8/29	9/29	1	26	0/26	14/26	<0.01
Sorghum	17	0/17	11/17	<0.01	24	1/24	9/24	<0.05	23	1/23	13/23	<0.01
Maize	24	0/24	5/24	<0.05	20	0/20	5/20	<0.05	25	0/25	14/25	<0.01

with low evolutionary rates tend to be retained after gene duplication (Davis and Petrov, 2004; Brunet et al., 2006). The high conservation of WGD duplicated genes might be caused by mechanisms of mutation buffering, where functionally important genes have more opportunities to buffer deleterious mutations when duplicated (Chapman et al., 2006). There are many other possible explanations; for example, duplicated genes are conserved mainly because they are central in biological networks, expressed at high level, or encode many protein domains. Thus, the conservation may be a by-product of other gene features (Gibson and Spring, 1998; Davis and Petrov, 2004; Gout et al., 2010). In this study, however, we found that the correlation of evolutionary rate and gene retention probability was still significant when the effects of many other gene features were excluded, including gene complexity, expression level, and GC3 content. It is more likely that evolutionary rate could affect the behavior of WGD duplicated genes alone. Sémon and Wolfe (2008) suggest that slowly evolving genes might be more prone to be subfunctionalized after gene duplication and, thus, more prone to be retained as duplicates. Indeed, our observation strongly supports such an important role of evolutionary rate. The role of structural complexity was also found to be consistent with previous studies, where it was reported that highly complex genes prefer to be retained after gene duplication (He and Zhang, 2005; Chapman et al., 2006). Nevertheless, gene or peptide lengths seem to have trivial contributions, while protein domains encoded by each gene or a gene's functional diversity more intensely affect evolutionary fates of duplicate genes after WGD.

Surprisingly, we observed that GC/GC3 content appeared as one of the most important contributors across all six plant genomes. The contribution of gene conversion was first used to explain the positive correlation between GC3 content and gene retention probability. It was reported that GC3 was elevated in duplicated genes that were subject to concerted evolution (Benovoy et al., 2005). However, later work found that the contribution of GC3 to gene conversion was limited (Sugino and Innan, 2006). Thus, other mechanisms might also underlie the correlation between GC3 content and the probability of gene retention. Recent studies have uncovered that genes of high GC3 content may provide more targets for DNA methylation in plants, leading to a high variability of gene expression (Tatarinova et al., 2010). As a result, it is likely that duplicated genes with high GC3 content also possess more targets for DNA methylation. Additionally, epigenetic mechanisms (e.g. DNA methylation) have long been recognized to play a significant role in rebuilding the transcription regulatory network after polyploidization (for review, see Doyle et al., 2008). Consequently, these duplicated genes are more prone to be subfunctionalized through methylation-mediated regulation reprogramming at the onset of WGDs, thus increasing the opportunity for survival after WGDs.

Direct comparison between retained and lost genes showed that genes with high expression levels were more prone to be retained after WGD, corresponding to previous studies, where duplicated genes expressed at higher dosages were more likely to be retained (Seoighe and Wolfe, 1999; Aury et al., 2006; Conant and Wolfe, 2007). However, expression level only exhibited a limited role in contributing to gene retention, as its effect was only observed in Arabidopsis and poplar after regression modeling. According to scenarios suggested by previous researchers, dosage increase for some kinds of genes might be beneficial to organism survival during evolution; thus, these duplicated genes expressed in higher dosages are more likely to be retained (Seoighe and Wolfe, 1999; Aury et al., 2006; Conant and Wolfe, 2007). However, for most genes in the genome, the dosage increase benefits might be weak and occasional. The increase of dosage might be neutral or even deleterious in most situations (Innan and Kondrashov, 2010). Therefore, it is not unexpected to observe that expression level plays a limited role in gene retention in this study.

All three major gene features we discovered here might be better explained by subfunctionalization models (Force et al., 1999). Genes with high complexity have diverse functions and thus serve as suitable targets for subfunctionalization (Force et al., 1999; He and Zhang, 2005). Genes with low evolutionary rate might also be more prone to be subfunctionalized. According to models proposed by Sémon and Wolfe (2008), slowly evolving genes might stay interchangeable during evolution for longer times and thus may have more opportunities to be subfunctionalized. For genes with high GC3 content, the intense methylation might quickly change their expression patterns and thus promote subfunctionalization at the expression level (Tatarinova et al., 2010). However, subfunctionalization, as an evolution model in general, could be applied to almost all modes of gene duplication, including both WGDs and SSDs. Comparisons between WGD retained and tandemly duplicated genes showed that gene features related to these two types of duplicated genes were distinct, indicating that the evolution of WGD duplicated genes may be unique. Indeed, this expectation is supported by previous empirical data proposing that WGD retained and tandemly duplicated genes shared quite different functional roles (Seoighe and Gehring, 2004; Maere et al., 2005). The uniqueness of the evolution of WGD retained genes could well be explained by the dosage balance model (Freeling and Thomas, 2006; Birchler and Veitia, 2007, 2010), which declares that the stoichiometric balance requirement among biological network members could force many genes to be retained after WGDs. Indeed, GO enrichment analysis in this study showed that duplicate genes retained after WGD were rarely duplicated by tandem duplication.

If all WGD retained genes evolved through subfunctionalization, we would expect that all GO categories enriched for these genes might also be enriched for tandemly duplicated genes. In this study, however, this pattern was only observed in a small fraction of GO categories (e.g. 14 out of 157 GO categories in Arabidopsis). Thus, the role of subfunctionalization might be trivial for WGD retained genes. In reverse, if all WGD retained genes evolved under dosage balance selection, then no GO categories enriched for WGD retained genes could also be enriched for tandemly duplicated genes. However, 14 of them were still observed to show this pattern. The seeming contradiction of these predictions derived from subfunctionalization and the dosage balance model may be explained by WGD retained genes being classified into three different types. In addition to the important role of neofunctionalization, the classification of retained genes suggests that both dosage balance selection and subfunctionalization might take place after WGDs. In general, type I genes were rarely duplicated by other modes of gene duplication (e.g. tandem duplications). It is evident that ohnologs belonging to this category always diverged slowly at both the sequence and expression levels. Low evolutionary rates were also observed not only between ohnologs but also among orthologous genes across different plant lineages. In addition, we observed that these genes are usually expressed at high levels and in most tissues. Additionally, genes belonging to type I are involved in many central biological processes. Thus, they were apparently selected by dosage balance. In contrast, type III genes could be duplicated through tandem duplication more often. They evolved rapidly both between ohnologs and orthologs; high expression divergence was observed between ohnologs, and a large number of ohnologs belonging to this category exhibited obvious patterns of asymmetric expression after WGDs. Most genes in this group were lowly expressed and only expressed in a few tissues. In addition, we observed that many genes belonging to type III were positively selected after WGDs. All the above-mentioned signatures suggest that type III genes act as targets for neofunctionalization. Interestingly, high GC/GC3 content was also found for type III genes. Thus, they may be more prone to be under DNA methylation-mediated epigenetic control at the onset of polyploidy (Shaked et al., 2001; Salmon et al., 2005; Lukens et al., 2006); therefore, such opportunities to survive might also increase through subfunctionalization immediately after WGDs. Different from retained genes belonging to either type I or type III, type II genes evolved in neither dynamic nor conserved manners. The main characteristic of type II genes was their high structural complexity. They encode the longest genes and proteins, contain many alternative splicing isoforms, and come with a large number of annotated protein domains. Thus, type II genes are good targets for subfunctionalization.

However, the evidence of subfunctionalization appeared limited for type II genes, at least at the expression level. We only found complementarily expressed patterns in sorghum and maize (Supplemental Table S11). Given that the expression data employed for sorghum and maize were measured by ESTs, the observed subfunctionalization of gene expression should be treated with caution, and further analyses are needed. In a recent study, analyses of vertebrate WGD events showed that about 24% of the ohnologs encoded proteins that differed in domain architecture and/or subcellular localization (Kassahn et al., 2009). The usefulness of this analysis would be helpful to detect signatures of subfunctionalization at the sequence level without the lack of functional data or protein structure in many of our analyzed genomes and data sets.

Although the above-mentioned results suggest that WGD retained genes might evolve under different evolutionary modes, the dosage balance model might serve as the most important one. It was found that approximately 2.3% to 9.4% of retained genes could also be duplicated through tandem duplications. Of them, type I genes were rarely tandemly duplicated (approximately 3.6%–6.6%; Table III), while type III genes were more prone to be tandemly duplicated (approximately 4.8%–9.4%; Table III). However, tandemly duplicated genes constituted approximately 15% to 25% of the whole gene set in the genome (data not shown). Thus, all types of WGD retained genes were indeed underrepresented for tandem gene duplications compared with the levels of genome averages, and the underrepresentation was further confirmed by functional data. Compared with GO categories enriched for both WGD retained genes and tandemly duplicated genes, many more GO categories enriched for WGD retained genes were underrepresented for tandem gene duplications (Supplemental Table S12). Results from gene feature comparisons could be explained by the dosage balance model as well. Highly complicated genes, especially genes with multiple protein domains, might serve as central nodes in biological networks; thus, their dosage changes might be easily selected. For genes involved in central cellular processes, such as ribosome-related genes, the dosage balance requirement is also strong. When all members of the network were duplicated after WGDs, the divergence rate between gene duplicates would also need to be slow, as high dosage balance was required. Even at the sequence level, the divergence between two duplicate copies was not preferable, as the change of sequence would alter the efficiency of a gene's original function. Thus, both type I and type II genes diverged more slowly at both the sequence and expression levels compared with type III genes. Type III genes were more often positively selected after WGD, and most of their ohnologs showed asymmetric expression patterns. Thus, they diverge and evolve quickly. However, we could not rule out the possibility that this type of gene was also selected by dosage balance. GO analysis showed limited evidence of overrepresentation for tandemly duplicated genes in type III, and this phenomenon was indeed similar for type I and type II.

In conclusion, this study strongly supports that the retention and succeeding evolution of duplicate genes derived from WGDs were intensely impacted by gene features, such as evolutionary rate, GC/GC3 content, and structural complexity. The integrated analyses of gene features classified the retained genes derived from WGDs into three distinct groups. Each group of retained genes was associated with unique gene features and evolutionary behaviors. Type I genes serve as good targets for balance gene drive selection. Indeed, we found that genes in this group were depleted for tandem gene duplications and evolved slowly, corresponding to the model hypothesized by balance gene drive (Freeling and Thomas, 2006). Type II genes might become targets for subfunctionalization, while type III genes were more prone to be neofunctionalized. Our results thus suggest that multiple mechanisms, including dosage balance, neofunctionalization, and subfunctionalization, may together drive the evolution of duplicated genes after WGDs. In all these mechanisms, dosage balance served as the most important selection force after WGDs. It appears likely that arguments between evolution models are primarily caused by tests built upon the random sampling of genes with diverse types of gene features.

MATERIALS AND METHODS

Genome Sequences and Annotation Data

Our study used six genomes representing diverse lineages of flowering plants (Supplemental Table S1). The genome annotation data of Arabidopsis (*Arabidopsis thaliana*; version 9.0) was downloaded from The Arabidopsis Information Resource (http://www.arabidopsis.org/); for rice (*Oryza sativa*), genome annotation data (version 7.0) maintained by Michigan State University (http://rice.plantbiology.msu.edu/) were used; maize (*Zea mays*) genome annotation data (release 4a.53) were downloaded from http://www.maizesequence.org/; the annotation data for the soybean (*Glycine max*) genome were downloaded from Phytozome (www.phytozome.net/soybean); and annotation data for the other two genomes, sorghum (*Sorghum bicolor*) and poplar (*Populus trichocarpa*), were downloaded from the Ensembl database (http://plants.ensembl.org/index.html; Hubbard et al., 2009).

Identification of the WGD Events

Before building duplication blocks, the annotation data of the six studied plant genomes were cleaned. transposable element-related sequences were removed from the data set. If a locus contained multiple alternative splice forms, only the longest one was retained. For tandem arrays of duplicated genes, only one sequence was selected as a representative for further analyses. Tandem duplications were detected through these methods as follows: if (1) two proteins could be aligned by BLAST (Altschul et al., 1997) with an E value of 1e-10 or less and identity of 30% or greater, (2) the aligned region covered more than 80% of the longer sequence, and (3) the genes were separated by fewer than 10 genes on the chromosome, they were considered a pair of tandemly duplicated genes. After data cleaning, the following procedures were used to build syntenic duplication blocks. First, putative anchor points of WGDs were identified by BLAST (Altschul et al., 1997). If two proteins were aligned with an E value of 1e-20 or less, the aligned region covered more than 50% of the longer sequence, and the E value did not exceed 1e-20 times the best non-self-hit's E value, they were retained as putative anchor points. However, if one query sequence fulfilled more than 20 hits of these BLAST parameters, both query and hit proteins were removed from the data set to exclude large gene families. Second, DAGchainer (Haas et al., 2004) was applied to characterize duplicated blocks in Arabidopsis and maize with the parameter "-s –I D=20 A=6." Third, in order to estimate the approximate origin times of these duplicated blocks, DS between pairs of anchor genes in each block were calculated. Because gene pairs from one sister region were almost simultaneously duplicated, the median DS value between these gene pairs could be used as an approximation to date the current synteny duplication block pairs. We combined the DS distribution box plots for each pair of duplication blocks into one huge graph, based on which syntenic duplication block pairs derived from different rounds of WGD could be distinguished without difficulty (Supplemental Fig. S1). Lastly, a manual check was carried out to exclude nesting or large overlapping blocks after grouping blocks into different rounds of WGD based on DS distribution patterns. Since the age distribution for WGD events differs broadly along different lineages, parameters used for each procedure may vary from one species to another. Given the case of ancient WGD events, such as the grass-shared WGD event that occurred about 70 million years ago (Paterson et al., 2004), we loosened the parameters with an E value adjusted to 1e-5.

Classification of Retained and Lost Genes

To identify gene features determining the retention probability of duplicate genes derived from WGDs, we classified them into two groups, called retained and lost genes. As the studied plant genomes underwent several rounds of WGDs, some genes might be considered as retained in the recent round of WGDs but lost during ancient WGD events. To distinguish retained and lost genes, we used a simple paradigm by collecting all syntenic blocks related to the WGD event under investigation, such as blocks derived from the α event in Arabidopsis (the most recent round of WGD; described by Bowers et al. [2003]). Then, duplicate genes with both copies present in syntenic blocks were classified as retained genes, while those without any corresponding duplicate copy in syntenic blocks were classified as lost genes. Although all genes were duplicated after WGD, we excluded genes that could not be assigned onto syntenic blocks, as their evolutionary status could not be clearly inferred.

The reconstruction of syntenic blocks related to ancient WGD events appears difficult and error prone. Thus, we merely focused on genes that came from the most recent round of WGDs (named R1 WGD), which is defined as groups of blocks that are youngest according to the DS distribution plot. In this study, blocks of R1 WGD could be straightforwardly dissected for all three dicot species and one grass species, maize (Supplemental Fig. S1). However, the other two grass species, rice and sorghum, contained several recent syntenic duplicate blocks in addition to the major age groups. These blocks were also treated as R1 blocks in this study, as they were found to have heavily undergone gene conversion (Paterson et al., 2009). The R1 blocks contained quite a few overlaps, which might be caused by small-scale segmental duplications or small fragments left over from ancient WGDs. Note that it could also come from DS overlaps if older homeologues were interspersed with younger ones. We resolved these overlapping blocks by removing both small fragments nested in large duplication blocks and small, ancient fragments overlapped by those large and young blocks. This data mining was again performed by adjusting parameters of DAGchainer (Haas et al., 2004). After following the above-mentioned steps, we collected a relatively clean data set containing syntenic blocks only from the R1 WGD (Fig. 1). However, a quantity of overlaps still existed and could not be excluded from R1 blocks, ranging from 0.15% to 2% in total for different genomes (Supplemental Table S2). Since we placed emphasis on gene comparisons, these tiny overlaps would not severely influence subsequent data analyses. Indeed, less than 0.1% of duplicate genes were covered by these tiny overlapping blocks (Supplemental Table S2).

Estimation of Sequence Divergence and Evolutionary Rates

In order to estimate the sequence divergence between duplicated gene pairs generated by WGD (ohnologs), protein sequences for these gene pairs were aligned by ClustalW (Thompson et al., 1994). Then, protein sequence alignments were transformed into coding sequence alignments by PAL2NAL (Suyama et al., 2006). Sequence divergence was further calculated by using codeml incorporated in PAML using the F3X4 model (Yang, 1997). The evolutionary rates of genes were calculated by ortholog comparisons between related genomes, for example, rates of evolution in Arabidopsis were estimated by comparing with their orthologs in *Arabidopsis lyrata* (for all genome comparisons used in this study, see Supplemental Table S1). Orthologous genes were downloaded from the Ensembl database (http://plants.ensembl.org/index.html) or Phytozome (www.phytozome.net) through a biomart portal. Only one-to-one orthologs were considered in most of the genomes except the maize and soybean genomes, due to recent WGD events in these two genomes after they diverged from their related genomes. For retained genes of maize and soybean, we estimated evolutionary rates for both duplicated copies and then calculated average values for the two copies for the purpose of further analysis. Rates of evolution, including DN, DS, and DN/DS ratio, were separately calculated between ortholog gene sequences using PAML as described above.

The extent of positive selection after WGD was also tested through PAML (http://abacus.gene.ucl.ac.uk/software/paml.html). The branch-site model M2b (assuming positive selected sites in foreground branches) was tested against M2a (assuming no positive selected sites in foreground branches). Then, the positively selected sites in all foreground branches were extracted from the tests results (Zhang et al., 2005). Here, ohnologs generated by one

WGD event were used as foreground branches, while their orthologs from another genome that did not contain the current WGD event were used as background branches. For example, the α event was considered to take place in Arabidopsis but not in poplar, so ohnologs generated from the α event could be compared with their orthologs in poplar. We carried out tests of positive selection only in the four species Arabidopsis, poplar, soybean, and maize. Rice and sorghum were not under investigation due to the difficulty in identifying a proper outgroup species.

Expression Data Analysis

The Arabidopsis microarray data were downloaded from AtGenExpress (http://www.weigelworld.org/research/projects/geneexpressionatlas). This data set was obtained by using an Affymetrix ATH1 chip to estimate gene expression for a total of 79 tissues and development time points (Schmid et al., 2005). The information for genes matching probe sets was downloaded from Affymetrix (http://www.affymetrix.com). Expression data for rice and poplar were downloaded from PLEXdb (http://www.plexdb.org/). For rice, experiment OS 32 was used, which used the Affymetrix rice genome array to estimate gene expression for a total of 39 tissues/organs (Wang et al., 2010). Probe consensus sequences were BLASTed against rice genes, and only probes that could be aligned with more than 50% of their lengths and higher than 95% identity to gene sequences were considered in this analysis. For poplar, experiment PT2 was used, and gene expression status for nine tissues was estimated by Affymetrix poplar genome array (Wilkins et al., 2009). Information for probe sets matching genes was also downloaded from Affymetrix (http://www.affymetrix.com). All microarray data were normalized using the robust multiarray average method (Irizarry et al., 2003). Then, the present call information output by MAS was used to determine whether one probe was expressed (Gautier et al., 2004). Probes matching multiple genes were excluded for further analyses in all three arrays used in this study. If one gene corresponds to multiple probes, the average intensity between probes was used. Since we failed to find genome-wide microarray data for maize and sorghum in public domains, UniGene EST profiles were used as a replacement. All coding sequences in these two genomes were searched against UniGene EST data sets, and the best hits were considered as representatives of these genes. To exclude suspicious gene mapping, only those BLAST pairs with at least 95% sequence identity between them were retained. The expression level for one specific gene was estimated by averaging its expression values over all tissues. Finally, expression data for soybean were downloaded from the transcriptome atlas of the soybean Web site (http://digbio.missouri.edu/soybean_atlas/). In this data set, high-throughput transcriptome sequencing techniques (Solexa) were used to get gene expression estimates from 14 tissues/conditions in soybean (Severin et al., 2010). For all of the expression data used here, the level of average expression for each gene between different tissues/conditions was calculated, and expression specificity (τ) for each gene was estimated according to the formula below:

$$\tau = \sum_{j=1}^{n} 1 - \left(\frac{\log s(j)}{\log s_{\max}}\right) \Big/ (n-1)$$

where $s(j)$ denotes the expression value in each tissue, $S_{\max}$ denotes the maximum expression value in all tissues, and n is the number of tissues. τ ranges from 0 to 1, with values close to 0 indicating widespread expressed genes and values close to 1 indicative of biased expressed genes. When $\tau = 1$, this gene is expressed in only one tissue (Yanai et al., 2005).

To examine the pattern of expression divergence between duplicated genes, the Pearson's correlation coefficient was calculated. In addition, we used a simple rule to identify ohnologs with similar or asymmetric expression patterns according to their spatial expression domains. When both copies were expressed in more than 70% of the tissues, they were considered the same. Meanwhile, if one copy was expressed in more than 70% of the tissues while the other was expressed in less than 30%, these two copies were considered asymmetric. All remaining ohnologs were considered to be complementarily expressed. Ohnologs with one of their copies missing expression signals in all tissues were excluded from further analyses.

Statistical Analysis

To explore whether the retention and loss of duplicate genes after WGDs is associated with biological features, we compared several biological features between retained and lost genes. Features used in this study included gene length (genomic sequence length of genes), peptide length, intron total length, intron number, intron average length (intron length divided by intron number), with or without introns, protein domain diversity (number of different InterPro domains in each gene), with or without protein domain annotations, average expression level, expression specificity, DS, DN, selection constraints, number of transcript isoforms, number of alternative splicing isoforms (only genes with alternative splicing were considered), with or without alternative splicing isoforms, GC content, and GC3 content (GC content at the third codon position). We used Wilcox's rank-sum test to measure the effects of continuous variables. For discrete variables, Fisher's exact test was used.

Considering that gene features relevant to the retention probability of duplicated genes were highly correlated with each other, we carried out logistic regression modeling based on PCA to simplify questions under investigation and exclude problems of parameter correlation. Highly correlated biological features were first subjected to PCA, including gene length, peptide length, total intron length, average intron length, intron number, DN, DN/DS, expression level, and expression specificity. PCA components with eigenvalues greater than 1 were extracted. These extracted components and the remaining variables (DS, AS, domain number, GC content, and GC3 content) were submitted for logistic regression analyses. Each gene model was given an indicator variable of 0 or 1: 0 denoted lost genes while 1 denoted retained genes. All gene features were then tested against this new variable through logistic regression.

Classification of WGD Retained Genes

Retained genes after WGDs were classified based on their gene features. Here, features selected for gene classification include gene length, peptide length, intron length, intron average length, intron number, protein domain diversity, number of alternative splicing isoforms, average expression level, expression specificity, GC content, GC3 content, DN, DN/DS, and DS. All features were Z transformed before cluster analysis was performed. Then, the k-means cluster method was applied to classify these genes into different groups. The k-means clustering method is a standard partitioning clustering algorithm based on K centroids of a random initial partition that is iteratively improved (Hartigan and Wong, 1979). The algorithm is initiated by randomly partitioning the genes into k groups. Each group is then represented by a "centroid," and the genes are repartitioned to the cluster whose centroid is most similar to them. The partitioning process is iterated until the partitions are stable, attempting to maximize the similarity of the genes in each cluster. The end result of the algorithm is a set of k clusters of genes with similar features so that the means across clusters (for all variables) are as different from each other as possible.

GO Enrichment Analysis

GO (Ashburner et al., 2000) categories with more than 50 genes were tested for the enrichment of duplicate genes using the program package FUNC (Prufer et al., 2007). In this package, a hypergeometric test was used to identify GO categories with overrepresentation or underrepresentation of WGD retained genes and tandemly duplicated genes. To avoid overcounting of significant GO categories due to the hierarchical structure of GO annotation, each GO category was tested by subtracting genes belonging to all its child categories. The test outputs with $P < 0.01$ and a false discovery rate smaller than 0.01 were considered as significant.

Supplemental Data

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

Supplemental Figure S1. Distribution of DS between segmental duplication blocks.

Supplemental Figure S2. Genes retained after WGD were distinct from tandemly duplicated genes on the possibility of being alternatively spliced.

Supplemental Figure S3. Gene feature comparisons between WGD-retained genes and tandemly duplicated genes.

Supplemental Figure S4. Correlation matrix for all available gene features investigated in Arabidopsis.

Supplemental Figure S5. Pairwise DS differences between different types of WGD-retained ohnologs.

Supplemental Figure S6. Pairwise DN/DS differences between different types of WGD-retained ohnologs.

Supplemental Figure S7. Pairwise differences of expression correlation coefficient between different types of WGD-retained ohnologs.

Supplemental Table S1. The plant genomes used for the calculation of evolutionary rate in this study.

Supplemental Table S2. Genome coverage and homologous gene pairs derived from the most recent round of WGDs in the six plant genomes.

Supplemental Table S3. Comparison of gene features between WGD-retained and WGD-lost genes.

Supplemental Table S4. PCA for the highly correlated gene features.

Supplemental Table S5. Clustering results for WGD-retained genes in the Arabidopsis genome.

Supplemental Table S6. Clustering results for WGD-retained genes in the poplar genome.

Supplemental Table S7. Clustering results for WGD-retained genes in the soybean genome.

Supplemental Table S8. Clustering results for WGD-retained genes in the rice genome.

Supplemental Table S9. Clustering results for WGD-retained genes in the sorghum genome.

Supplemental Table S10. Clustering results for WGD-retained genes in the maize genome.

Supplemental Table S11. Tissue expression pattern between pairs of duplicated genes for different types of WGD-retained genes.

Supplemental Table S12. Enrichment for functional classes in different types of WGD-retained homologs in Arabidopsis.

Supplemental Table S13. GO categories shared between or unique to each type of WGD-retained genes in Arabidopsis.

Received May 23, 2012; accepted January 31, 2013; published February 8, 2013.

LITERATURE CITED

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:** 3389–3402

Amoutzias GD, He Y, Gordon J, Mossialos D, Oliver SG, Van de Peer Y (2010) Posttranslational regulation impacts the fate of duplicated genes. Proc Natl Acad Sci USA **107:** 2967–2971

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al (2000) Gene Ontology: tool for the unification of biology. Nat Genet **25:** 25–29

Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Ségurens B, Daubin V, Anthouard V, Aiach N, et al (2006) Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature **444:** 171–178

Barker MS, Kane NC, Matvienko M, Kozik A, Michelmore RW, Knapp SJ, Rieseberg LH (2008) Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. Mol Biol Evol **25:** 2445–2455

Benovoy D, Morris RT, Morin A, Drouin G (2005) Ectopic gene conversions increase the G + C content of duplicated yeast and Arabidopsis genes. Mol Biol Evol **22:** 1865–1868

Birchler JA, Veitia RA (2007) The gene balance hypothesis: from classical genetics to modern genomics. Plant Cell **19:** 395–402

Birchler JA, Veitia RA (2010) The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. New Phytol **186:** 54–62

Blanc G, Barakat A, Guyot R, Cooke R, Delseny M (2000) Extensive duplication and reshuffling in the *Arabidopsis* genome. Plant Cell **12:** 1093–1101

Blanc G, Hokamp K, Wolfe KH (2003) A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. Genome Res **13:** 137–144

Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. Plant Cell **16:** 1679–1691

Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature **422:** 433–438

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. Mol Biol Evol **23:** 1808–1816

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 USA **103:** 2730–2735

Conant GC, Wolfe KH (2007) Increased glycolytic flux as an outcome of whole-genome duplication in yeast. Mol Syst Biol **3:** 129

Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A, et al (2006) Widespread genome duplications throughout the history of flowering plants. Genome Res **16:** 738–749

Davis JC, Petrov DA (2004) Preferential duplication of conserved proteins in eukaryotic genomes. PLoS Biol **2:** E55

Davis JC, Petrov DA (2005) Do disparate mechanisms of duplication add similar genes to the genome? Trends Genet **21:** 548–551

Dehal P, Boore JL (2005) Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol **3:** e314

Doyle JJ, Flagel LE, Paterson AH, Rapp RA, Soltis DE, Soltis PS, Wendel JF (2008) Evolutionary genetics of genome merger and doubling in plants. Annu Rev Genet **42:** 443–461

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151:** 1531–1545

Freeling M (2008) The evolutionary position of subfunctionalization, downgraded. Genome Dyn **4:** 25–40

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

Freeling M, Thomas BC (2006) Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res **16:** 805–814

Gao LZ, Innan H (2004) Very low gene duplication rate in the yeast genome. Science **306:** 1367–1370

Gautier L, Cope L, Bolstad BM, Irizarry RA (2004) affy: analysis of Affymetrix GeneChip data at the probe level. Bioinformatics **20:** 307–315

Gibson TJ, Spring J (1998) Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. Trends Genet **14:** 46–49, discussion 49–50

Gout JF, Kahn D, Duret L (2010) The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. PLoS Genet **6:** e1000944

Guan Y, Dunham MJ, Troyanskaya OG (2007) Functional analysis of gene duplications in Saccharomyces cerevisiae. Genetics **175:** 933–943

Haas BJ, Delcher AL, Wortman JR, Salzberg SL (2004) DAGchainer: a tool for mining segmental genome duplications and synteny. Bioinformatics **20:** 3643–3646

Hakes L, Pinney JW, Lovell SC, Oliver SG, Robertson DL (2007) All duplicates are not equal: the difference between small-scale and genome duplication. Genome Biol **8:** R209

Hartigan J, Wong M (1979) Algorithm AS136: a k-means clustering algorithm. Appl Stat **28:** 100–108

He X, Zhang J (2005) Gene complexity and gene duplicability. Curr Biol **15:** 1016–1021

He X, Zhang J (2006) Higher duplicability of less important genes in yeast genomes. Mol Biol Evol **23:** 144–151

Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, et al (2009) Ensembl 2009. Nucleic Acids Res **37:** D690–D697

Innan H, Kondrashov F (2010) The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet **11:** 97–108

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics **4:** 249–264

Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, et al (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature **431:** 946–957

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, et al (2011) Ancestral polyploidy in seed plants and angiosperms. Nature **473:** 97–100

Jin L, Kryukov K, Clemente JC, Komiyama T, Suzuki Y, Imanishi T, Ikeo K, Gojobori T (2008) The evolutionary relationship between gene duplication and alternative splicing. Gene **427:** 19–31

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

Kellis M, Birren BW, Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature **428:** 617–624

Kimura M, King JL (1979) Fixation of a deleterious allele at one of two "duplicate" loci by mutation pressure and random drift. Proc Natl Acad Sci USA **76:** 2858–2861

Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV (2002) Selection in the evolution of gene duplications. Genome Biol **3:** RESEARCH0008

Kopelman NM, Lancet D, Yanai I (2005) Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. Nat Genet **37:** 588–589

Li L, Huang Y, Xia X, Sun Z (2006) Preferential duplication in the sparse part of yeast protein interaction network. Mol Biol Evol **23:** 2467–2473

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

Liang H, Plazonic KR, Chen J, Li WH, Fernández A (2008) Protein underwrapping causes dosage sensitivity and decreases gene duplicability. PLoS Genet **4:** e11

Lukens LN, Pires JC, Leon E, Vogelzang R, Oslach L, Osborn T (2006) Patterns of sequence loss and cytosine methylation within a population of newly resynthesized *Brassica napus* allopolyploids. Plant Physiol **140:** 336–348

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

Lynch M, Conery JS (2003) The evolutionary demography of duplicate genes. J Struct Funct Genomics **3:** 35–44

Lynch M, Force A (2000) The probability of duplicate gene preservation by subfunctionalization. Genetics **154:** 459–473

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y (2005) Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci USA **102:** 5454–5459

Ohno S (1970) Evolution by Gene Duplication. Springer-Verlag, Berlin

Otto SP, Whitton J (2000) Polyploid incidence and evolution. Annu Rev Genet **34:** 401–437

Pál C, Papp B, Lercher MJ (2006) An integrated view of protein evolution. Nat Rev Genet **7:** 337–348

Papp B, Pál C, Hurst LD (2003) Dosage sensitivity and the evolution of gene families in yeast. Nature **424:** 194–197

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, et al (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature **457:** 551–556

Paterson AH, Bowers JE, Burow MD, Draye X, Elsik CG, Jiang CX, Katsar CS, Lan TH, Lin YR, Ming R, et al (2000) Comparative genomics of plant chromosomes. Plant Cell **12:** 1523–1540

Paterson AH, Bowers JE, Chapman BA (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proc Natl Acad Sci USA **101:** 9903–9908

Paterson AH, Chapman BA, Kissinger JC, Bowers JE, Feltus FA, Estill JC (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

Prachumwat A, Li WH (2006) Protein function, connectivity, and duplicability in yeast. Mol Biol Evol **23:** 30–39

Prufer K, Muetzel B, Do HH, Weiss G, Khaitovich P, Rahm E, Paabo S, Lachmann M, Enard W (2007) FUNC: a package for detecting significant associations between gene sets and ontological annotations. BMC Bioinformatics **8:** 41

Rodgers-Melnick E, Mane SP, Dharmawardhana P, Slavov GT, Crasta OR, Strauss SH, Brunner AM, Difazio SP (2012) Contrasting patterns of evolution following whole genome versus tandem duplication events in *Populus*. Genome Res **22:** 95–105

Salmon A, Ainouche ML, Wendel JF (2005) Genetic and epigenetic consequences of recent hybridization and polyploidy in *Spartina* (Poaceae). Mol Ecol **14:** 1163–1175

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Schölkopf B, Weigel D, Lohmann JU (2005) A gene expression map of *Arabidopsis thaliana* development. Nat Genet **37:** 501–506

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, et al (2010) Genome sequence of the palaeopolyploid soybean. Nature **463:** 178–183

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science **326:** 1112–1115

Sémon M, Wolfe KH (2007) Consequences of genome duplication. Curr Opin Genet Dev **17:** 505–512

Sémon M, Wolfe KH (2008) Preferential subfunctionalization of slow-evolving genes after allopolyploidization in *Xenopus laevis*. Proc Natl Acad Sci USA **105:** 8333–8338

Seoighe C, Gehring C (2004) Genome duplication led to highly selective expansion of the *Arabidopsis thaliana* proteome. Trends Genet **20:** 461–464

Seoighe C, Wolfe KH (1999) Yeast genome evolution in the post-genome era. Curr Opin Microbiol **2:** 548–554

Severin AJ, Woody JL, Bolon YT, Joseph B, Diers BW, Farmer AD, Muehlbauer GJ, Nelson RT, Grant D, Specht JE, et al (2010) RNA-Seq atlas of *Glycine max*: a guide to the soybean transcriptome. BMC Plant Biol **10:** 160

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

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

Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng CF, Sankoff D, Depamphilis CW, Wall PK, Soltis PS (2009) Polyploidy and angiosperm diversification. Am J Bot **96:** 336–348

Su ZX, Wang J, Yu J, Huang XQ, Gu X (2006) Evolution of alternative splicing after gene duplication. Genome Res **16:** 182–189

Sugino RP, Innan H (2006) Selection for more of the same product as a force to enhance concerted evolution of duplicated genes. Trends Genet **22:** 642–644

Suyama M, Torrents D, Bork P (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res **34:** W609–W612

Swigonová Z, Lai J, Ma J, Ramakrishna W, Llaca V, Bennetzen JL, Messing J (2004) Close split of sorghum and maize genome progenitors. Genome Res **14:** 1916–1923

Talavera D, Vogel C, Orozco M, Teichmann SA, de la Cruz X (2007) The (in)dependence of alternative splicing and gene duplication. PLoS Comput Biol **3:** 375–388

Tang H, Bowers JE, Wang X, Paterson AH (2010) Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. Proc Natl Acad Sci USA **107:** 472–477

Tatarinova TV, Alexandrov NN, Bouck JB, Feldmann KA (2010) GC3 biology in corn, rice, sorghum and other grasses. BMC Genomics **11:** 308

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

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, et al (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science **313:** 1596–1604

Vision TJ, Brown DG, Tanksley SD (2000) The origins of genomic duplications in Arabidopsis. Science **290:** 2114–2117

Walsh JB (1995) How often do duplicated genes evolve new functions? Genetics **139:** 421–428

Wang L, Xie W, Chen Y, Tang W, Yang J, Ye R, Liu L, Lin Y, Xu C, Xiao J, et al (2010) A dynamic gene expression atlas covering the entire life cycle of rice. Plant J **61:** 752–766

Wendel JF (2000) Genome evolution in polyploids. Plant Mol Biol **42:** 225–249

Wilkins O, Nahal H, Foong J, Provart NJ, Campbell MM (2009) Expansion and diversification of the *Populus* R2R3-MYB family of transcription factors. Plant Physiol **149:** 981–993

Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH (2009) The frequency of polyploid speciation in vascular plants. Proc Natl Acad Sci USA **106:** 13875–13879

Wu X, Qi X (2010) Genes encoding hub and bottleneck enzymes of the Arabidopsis metabolic network preferentially retain homeologs through whole genome duplication. BMC Evol Biol **10:** 145

Wu Y, Zhu Z, Ma L, Chen M (2008) The preferential retention of starch synthesis genes reveals the impact of whole-genome duplication on grass evolution. Mol Biol Evol **25:** 1003–1006

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

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci **13:** 555–556

Zhang JZ, Nielsen R, Yang ZH (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol **22:** 2472–2479