Evolutionary Dynamics and Functional Specialization of Plant Paralogs Formed by Whole and Small-Scale Genome Duplications

Lorenzo Carretero-Paulet[1] and Mario A. Fares*[,1,2]

[1]Instituto de Biologia Molecular y Celular de Plantas-IBMCP (CSIC-UPV), Integrative Systems Biology Group, Valencia, Spain

[2]Department of Genetics, School of Genetics and Microbiology, University of Dublin, Trinity College, Dublin, Ireland

***Corresponding author:** E-mail: mfares@ibmcp.upv.es.

Associate editor: Michael Purugganan

Abstract

Gene duplicates are a major source of evolutionary novelties in the form of new or specialized functions and play a key role in speciation. Gene duplicates are generated through whole genome duplications (WGD) or small-scale genome duplications (SSD). Although WGD preserves the stoichiometric relationships between duplicates, those arising from SSD are usually unbalanced and are expected to follow different evolutionary dynamics than those formed by WGD. To dissect the role of the mechanism of duplication in these differential dynamics and determine whether this role was shared across species, we performed a genome wide evolutionary analysis of gene duplications arising from the most recent WGD events and contemporary episodes of SSD in four model species representing distinct plant evolutionary lineages. We found an excess of relaxed purifying selection after duplication in SSD paralogs compared with WGD, most of which may have been the result of functional divergence events between gene copies as estimated by measures of genetic distances. These differences were significant in three angiosperm genomes but not in the moss species *Physcomitrella patens*. Although the comparison of models of evolution does not attribute a relevant role to the mechanism of duplication in the evolution duplicates, distribution of retained genes among Gene Ontology functional categories support the conclusion that evolution of gene duplicates depends on its origin of duplication (WGD and SSD) but, most importantly, on the species. Similar lineage-specific biases were also observed in protein network connectivity, translational efficiency, and selective constraints acting on synonymous codon usage. Although the mechanism of duplication may determine gene retention, our results attribute a dominant role to the species in determining the ultimate pattern of duplicate gene retention and reveal an unanticipated complexity in the evolutionary dynamics and functional specialization of duplicated genes in plants.

Key words: gene duplication, functional specialization, whole genome duplication, small-scale genome duplication.

Introduction

Gene duplication is a major source of evolutionary novelties in the form of new genes and gene functions performing a key role in generating phenotypic diversity and speciation (Lynch and Conery 2000). Gene duplicates can be generated through large-scale genome duplications, involving entire genomes [whole genome duplication (WGD)], or by more restricted events of duplication of small-scale genomic regions, involving one to a few genes [small-scale genome duplication (SSD)] (Lynch 2007).

Compared with other eukaryotes, WGD and SSD have been particularly frequent in plants (Blanc and Wolfe 2004b; Cui et al. 2006). For instance, it has been proposed that the genome of the plant model species *Arabidopsis thaliana* has undergone at least three rounds of duplication during its evolutionary history (Vision et al. 2000; Simillion et al. 2002; Blanc et al. 2003; Bowers et al. 2003). Different surveys revealed that more than one-half of the genes in the *A. thaliana* genome are duplicates (Arabidopsis Genome Initiative 2000; Lynch and Conery 2000). Furthermore, polyploidy formed by WGD or hybridization of plants species is believed to have played an important role in the explosive evolutionary diversification of angiosperms (De Bodt et al. 2005), although polyploidy has also been observed in basal plant species (Rensing et al. 2007).

Because of their role in generating functional innovation and evolutionary diversification, evolution of duplicated genes has received much attention in the last decade. Early theoretical models predicted that, in most cases, one duplicated gene retains the ancestral function while the other one evolves neutrally, free from selective constraints, becoming inactivated due to the stochastic accumulation of deleterious mutations or even deleted from the genome (nonfunctionalization), with a small fraction of duplicates being retained after fixing gain-of-function mutations (neofunctionalization) (Ohno 1970; Zhang 2003; Moore and Purugganan 2005). Against this prediction, the fraction of duplicates retained was found to be larger than anticipated by theory (Zhang 2003).

Plausible explanations for the high levels of gene retention have been put forward: 1) gene duplication endows

© The Author 2012. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Research article

Downloaded from http://mbe.oxfordjournals.org/ at CSIC on March 1, 2013

organisms with mutational robustness as a result of functional redundancy (Gu et al. 2003); 2) selection for increased gene dosage (Conant and Wolfe 2008); and 3) opportunity for functional specialization. Two different models have been proposed to explain functional specialization (sub-functionalization): partitioning of the ancestral functions by the Duplication, Degeneration, Complementation model (Force et al. 1999) and optimization of duplicated genes for different ancestral secondary sub-functions by the Escape from Adaptive Conflict model (Conant and Wolfe 2008). These models also set the ground to explain how novel functions (neo-functionalization) may emerge from gene copies initially retaining sub-functions (He and Zhang 2005) or dosage selection (Francino 2005). In the light of population genetics, both relaxed purifying selection and positive Darwinian selection may drive the retention of duplicated genes (Zhang 2003; Conant and Wolfe 2008). The relative contribution of the evolutionary forces in the preservation of gene duplicates remains, however, a major question in molecular evolution (Zhang 2003; Conant and Wolfe 2008; Innan and Kondrashov 2010).

Some authors have proposed that gene duplicates resulting from WGD are more likely to be preserved than those arising from SSD, as the stoichiometry of WGD duplicated gene products remains balanced (Freeling and Thomas 2006; Hakes et al. 2007). Conversely, SSD is likely to yield gene copies that upset dosage balance, hence resolving in the nonfunctionalization of one of the gene copies (Lynch and Conery 2000). Those SSD duplicates that are not constrained by dosage balance can more readily undergo functional divergence. This divergence is more likely to occur in proteins with promiscuous functions, such as certain enzymes (Aharoni et al. 2005). Additional factors, including population size, generation time, frequency of recombination between duplicates and the mechanism of gene duplication itself also have a key role in the fate of duplicates (Lynch et al. 2001).

Previous studies in *A. thaliana* showed sharp differences in the retention dynamics of WGD and SSD duplicates from different functional classes (Blanc and Wolfe 2004a; Maere et al. 2005). WGD duplicates present both broader and higher expression levels than SSD duplicates (Ganko et al. 2007) and slower evolutionary rates (Yang and Gaut 2011). However, the extent to which these observations are true in all plant species is unknown. Because retained genes and functions ultimately define the phenotype of the organism and its ability to interact with the environment, we hypothesize that species-specific selective pressures play an important role in determining the contribution of WGD and SSD to the retention and specialization of gene duplicates.

To test this hypothesis, we investigated the distribution of evolutionary rates, codon-based evolutionary models and patterns of functional diversification in contemporary duplicates formed by WGD and SSD in four plant species, namely *A. thaliana*, *Populus trichocarpa*, *Zea mays*, and the species *Physcomitrella patens*. These four species represent three main plant evolutionary lineages, including dicots (*A. thaliana* and *P. trichocarpa*), monocots (*Z. mays*), or basal land plants (*P. patens*); develop different life forms, herbaceous (*A. thaliana*, *Z. mays*, and *P. patens*) or woody (*P. trichocarpa*); display distinct modes of reproduction, self-fertilization (*A. thaliana*), partial cross-pollination (*Z. mays*), or entirely cross-pollination (*P. trichocarpa*); and show prevalence of the haploid gametophyte (*P. patens*) or the diploid sporophyte (*A. thaliana*, *P. trichocarpa* and *Z. mays*) during life cycle. Our results reveal unprecedented patterns of evolution and retention of duplicates that highlight the influence of the species in the differential retention and specialization of duplicates after WGD and SSD.

Materials and Methods

Identification of Triplets of Sequences Formed by Paralogous Gene Pairs and a Close Ortholog

Protein and coding sequences (together with their corresponding annotations) from eight whole-genome sequenced plant species were downloaded from PLAZA v2.0 (http://bioinformatics.psb.ugent.be/plaza/) (Proost et al. 2009). Paralogous pairs of duplicated genes were defined as the resulting best reciprocal hits from all-against-all BLAST-searches using BLASTP with an *E*-value cutoff of 1*E*−5 and a bit score cutoff of 50 (Altschul et al. 1997). Prior to the analysis, we discarded: 1) sequences with significant similarity to transposable elements (*E*-values $<$ 1*E*−15 in BLASTN searches against sequences of the RepBase v16.03 database (Jurka et al. 2005); 2) tandem duplicates (sequences residing $<$15 genes or 100 kb apart on the chromosome); and 3) sequences alignable over a length of $<$150 amino acids or showing an identity score $<$30% (Li et al. 2001), according to ClustalW alignments of protein sequences (Thompson et al. 1994).

As synonymous substitutions do not result in amino acid replacements, it is assumed that they accumulate changes in a neutral manner, reflecting the overall mutation rate for that species. Therefore, the number of nucleotide substitutions per synonymous site (Ks) may be used as a good proxy of the divergence time between paralogs (Maere et al. 2005). Ks values of all paralogous genes were estimated on the basis of alignments of codon sequences obtained through Muscle (Edgar 2004) using the corresponding protein sequences as alignment guides. Codon alignments were further edited to eliminate poorly aligned positions and divergent regions using Gblocks (Castresana 2000). Pairwise Ks, the number of nucleotide substitutions per nonsynonymous site (Ka) and the Ka/Ks (ω) rates ratio were estimated using the maximum likelihood method implemented in the *codeml* program (Goldman and Yang 1994) of the PAML package v4.4 (Yang 2007). All duplicates were classified according to their age estimated by Ks values, according to the bibliography, and those inferred to have originated at the same time of the youngest WGD event in each species under study were retained (fig. 1): *A. thaliana*, 24–40 Ma ($0.72 \leq Ks \leq 0.99$) (Blanc et al. 2003); *Physcomitrella patens*, 30–60 Ma ($0.6 \leq Ks \leq 1.1$) (Rensing et al. 2007); *P. trichocarpa*, 8–13 Ma ($0.2 \leq Ks \leq 0.3$) (Sterck et al. 2005); and *Z. mays*, 5–12 Ma ($0.1 \leq Ks \leq 0.3$) (Blanc and Wolfe 2004b). Subsequently, paralogous genes were classified as resulting from WGD or SSD, based on their mapping to recognizable duplicated genomic segments

Downloaded from http://mbe.oxfordjournals.org/ at CSIC on March 1, 2013

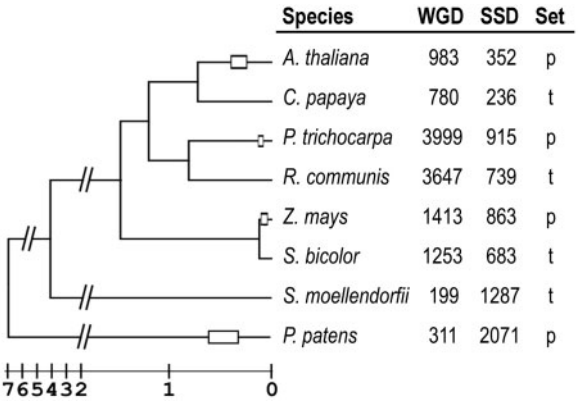

Fig. 1. Evolutionary relationships of plant species. A phylogenetic tree depicting evolutionary relationships among eight plant species examined is shown drawn to scale. The scale bar represents 100 My and was compressed below 200 My for clarity. The most recent WGD duplication events are represented as rectangles over the corresponding branches, with widths proportional to current date estimates. The numbers of pairs of paralog sequences from the ingroup species (p) and triplets of sequences formed by paralogs plus an ortholog from the outgroup species (t) examined here are shown on the right side table.

remaining from genome duplications according to the anchorpoints provided by PLAZA (Proost et al. 2009). Finally, we obtained triplets by adding to the pair of gene copies an orthologous sequence from a species predating the genome duplication event: *A. thaliana* (*Carica papaya*), *P. patens* (*Selaginella moellendorfii*), *P. trichocarpa* (*Ricinus communis*), and *Z. mays* (*Sorghum bicolor*) (fig. 1). Orthologs were defined as the resulting best reciprocal hits from all-against-all BLAST-searches using BLASTP with an *E*-value cutoff of 1*E*−5 and a bit score cutoff of 50.

Evolutionary Analysis

To characterize the molecular evolutionary forces involved in the retention of paralogous genes after duplication, Gblocks-edited codon alignments of triplets of sequences and manually constructed input trees were analyzed by two different classes of models of evolution as implemented in the program codeml (Yang 2007): the branch-specific and the branch-site–specific models. The branch-specific models estimate the nonsynonymous-to-synonymous substitution rates ratio (ω = Ka/Ks) for each of the branches of a tree, allowing us to test specific models of evolution along selected (foreground) paralogous branches in the tree. In the case of triplets (two paralogs and an ortholog), we tested for asymmetric sequence evolution, that is, whether one of the gene copies evolved at a different rate than the other two (two-ratios branch Model 2) by comparing such a model to one in which all rates were evolving at the same rate (one-ratio Model M0). We also tested whether both of the gene duplicates were evolving at different rates from one another and compared with the ortholog sequence after duplication (free-ratio branch Model M1). There is no null model to allow testing explicitly neutral evolution acting on specific paralogous branches. Alternatively, we called test of "non-neutral" evolution to one comparing the log-likelihoods between the two-ratio branch Model 2 with ω fixed at one for the selected branch and the two-ratios branch Model 2 with ω estimated from the data.

The branch-site models allow ω to vary in selected branches on the tree and also across codons in the sequences by defining different ω ratios site-classes (Yang and Nielsen 2002). Among them, clade Model D, allowing a class of sites to be under different substitution rates among the selected branch and the rest of the tree (Bielawski and Yang 2004), and Model A (Zhang et al. 2005), featuring an extra class of sites with $\omega > 1$ in foreground branches to be estimated from the data, were implemented. The comparison between the site-specific discrete null Model 3 (which allows the ω ratio to vary among sites, but holding ω constant among branches in the tree) and the clade Model D is an useful test to measure differential selective pressures acting on a significant number of amino acids (codon sites) either under relaxed purifying selection or under positive selection (PS) (Bielawski and Yang 2004). Finally, the comparison between the Model A with ω fixed at one for the examined branch as null model and the Model A was used as a conservative test to detect PS as opposed to relaxed purifying selection affecting a few sites in the selected branch (Zhang et al. 2005).

The comparison of models was performed through Likelihood Ratio Tests (LRT), which examine significance of differences between lnL resulting from two nested models (Goldman and Yang 1994). These test statistics are calculated as 2ΔLnL, twice the difference between the lnL under each model. If the null (simpler) model is true, LRT asymptotically follows a χ^2 distribution with a number of degrees of freedom equal to the differences in number of parameters between models. Hence, the LRT allows examining different evolutionary hypotheses by calculating a probability for the fitting of the examined data set to the alternative model being tested.

Finally, to test for natural selection acting on synonymous codon usage, we applied the mutation-selection codon substitution model, recently implemented in codeml, on the concatenated alignment of pairs of paralogous codon sequences (Yang and Nielsen 2008).

Measures of Genetic Distances between Paralogs

Genetic distance (d) between each paralog and the common ancestor was first measured using JTT amino acid distances (Jones et al. 1992) as

$$d_{\text{anc}-i} = \frac{d_{i-o} + d_{i-j} - d_{j-o}}{2}.$$

Here anc, *i*, *j*, and *o* are the ancestral node, paralogs *i* and *j*, and the ortholog, respectively. These distances were used to compare the divergence between gene copies for duplicates formed by SSD to those for duplicates originated from WGD. To remove the effect of the magnitude of genetic

Downloaded from http://mbe.oxfordjournals.org/ at CSIC on March 1, 2013

distance and the age of the duplication on the estimates, we normalized the differences between paralogs *i* and *j* as

$$\Theta = \frac{\mathrm{abs}\left(d_{\mathrm{anc}-i} - d_{\mathrm{anc}-j}\right)}{d_{\mathrm{anc}-i} + d_{\mathrm{anc}-j}}.$$

A. thaliana Protein–Protein Interactions Data

We used the protein–protein interaction (PPI) data set from the release 2.0 of the *A. thaliana* predicted interactome, available for downloading at TAIR (www.arabidopsis.org). This network consists of a set of 72,266 predicted interactions involving 7,177 proteins, of which about 5,134 interactions involving 2,617 unique proteins were experimentally confirmed (merging data sets from TAIR, IntAct-EBI and BIND/BOND). The prediction algorithm (Geisler-Lee et al. 2007) began with the identification of orthologs for *At* species in seven other species (*Escherichia coli*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Mus musculus*, and *Homo sapiens*), for which we had partial information on protein interactions. We considered two proteins in *A. thaliana* to interact if such an interaction was observed in more than one species.

Measures of Codon Adaptation Indexes

To estimate average translational rates we used Codon Adaptation Indexes (CAI). CAI values range from 0 to 1, with higher values indicating a higher proportion of the most frequently used synonymous codons. Codon usage was firstly determined for all genes in each species using the DAMBE software, and subsequently used as reference set of the frequencies of the codons of most expressed genes in each species. CAI for individual genes was calculated using the method depicted in (Xia 2007) as implemented in DAMBE. Differences in CAI between paralogs were normalized by the sum of CAIs so that all values could be comparable. For example, the difference between CAI for gene A and CAI for gene B was defined as $[\mathrm{abs}(\mathrm{CAI_A} - \mathrm{CAI_B})/(\mathrm{CAI_A} + \mathrm{CAI_B})]$.

Functional Categorization of WGD and SSD Duplicates

Paralogous genes were assigned to functional categories using their associated Gene Ontology (GO) terms as provided by the PLAZA functional annotation database (Ashburner et al. 2000; Proost et al. 2009). To have a broader overview of the ontology content, GO terms were remapped to the corresponding plant GO Slim terms using CateGOrizer (Zhi-Liang, Bao, Reecy 2008). Note that child terms in the ontology can have more than a parental GO Slim term. In addition, a gene might have more than one distinct function and therefore might be annotated with more than one GO term. As a result, the total number of functional classifications is greater than the total number of genes. We performed enrichment analysis of GO terms of each subset of gene duplicates by comparing with the full data set of genes for each genome using Fisher's exact test. Resulting *P* values were corrected to control for multiple testing by calculating the false discovery rate.

Results and Discussion

Higher Evolutionary Rates in SSD than in WGD Duplicates

We identified duplicates according to the methodology provided in Materials and Methods. We selected duplicates arising during the predicted time of occurrence of the most recent WGD in four plant species, according to nucleotide substitutions per synonymous site (Ks) rates. We classified duplicated genes as resulting from WGD or SSD based on their mapping to recognizable duplicated genomic segments remaining from genome duplications according to the anchorpoints provided by PLAZA. Thus, we obtained an accurate sampling of the duplicates generated by WGD and SSD from comparable evolutionary time points to dissect the role of the mechanism of duplication in gene retention (fig. 1). The estimated dates for the younger WGD (*Z. mays* and *P. trichocarpa*) ranged between 6 and 13 My, whereas for the older (*A. thaliana* and *P. patens*) were between 24 and 60 My (fig. 1).

Using this data set, we performed pairwise estimations of the number of nucleotide substitutions per synonymous site (Ks), nonsynonymous site (Ka), and ω rates ratio (Ka/Ks) in duplicates arising from the most recent WGD, and contemporary SSD, events in four plant genomes (fig. 2 and supplementary table S1, Supplementary Material online). Estimates of ω provide a simple and useful measurement of the strength of selection operating between duplicated gene copies (simplistically, $\omega = 1$, $\omega < 1$ and $\omega > 1$ indicates neutral, purifying, and PS, respectively). The means of the distributions of ω values ranged from 0.174 to 0.326, revealing that at least one gene copy has been evolving under purifying selection most of the evolutionary time (fig. 2*c*).

We observed a slight but significant increase in the rates of evolution (greater ω values) between SSD duplicates compared with WGD duplicates in all, but *P. patens*, species (fig. 2). Higher ω ratios in SSD duplicates were due to a higher accumulation of nonsynonymous substitutions (greater Ka values) and not to lower Ks values, indicating that WGD duplicates are under stronger selective constraints than those formed by SSD. In addition, the strength of natural purifying selection after gene duplication correlated with the estimated time of duplication (fig. 2*c*), despite possible biases owing to the expected variations in mutation rates between species. This correlation may be the result of the following: 1) the presence of slightly deleterious nonsynonymous mutations unfiltered by purifying selection in the younger gene duplications; 2) greater Ka values due to recent functional divergence events whose signatures on Ka/Ks values have not yet been diluted by high Ks values (assumed to be neutral, hence proportional to the time of divergence); or 3) both these possibilities. Persistence of slightly deleterious mutations for longer periods is likely in populations with small effective sizes, such as in plants (Ohta 1973). This is unlikely to bias our results when comparing the patterns of evolution after gene duplication of SSD to WGD because these patterns were compared within each of the

Downloaded from http://mbe.oxfordjournals.org/ at CSIC on March 1, 2013

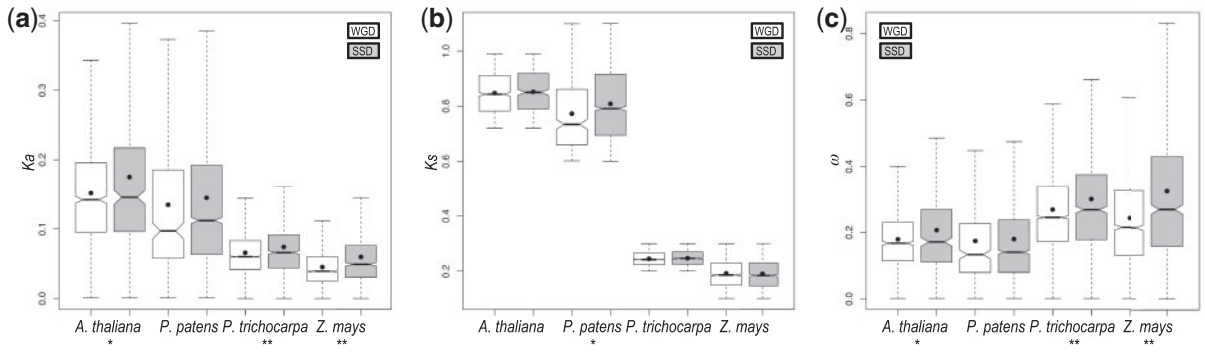

FIG. 2. Notched box-plots representation of pairwise estimates of Ka, Ks, and ω between paralogs. Black dots indicate means. Their differences were tested using a Student's t test and the resulting probabilities (P values *<1E−3, **<1E−6) are shown.

lineages, hence effects should have equal weights on both data sets (WGD and SSD). Indeed, the patterns of evolution of duplicates belonging to one or another mechanism of duplication appear to be insensitive to the time of duplication.

Differential Distribution of Evolutionary Models in WGD and SSD Duplicates among Species

The differences in the rates of evolution between duplicates from SSD and WGD (fig. 2) may reflect differences in the selection constraints between paralogs after duplication. To test whether the different modes of duplication (WGD and SSD) has been followed by differential evolutionary dynamics, we tested different evolutionary models using alignments of triplets of codon sequences under a maximum-likelihood framework (supplementary table S2, Supplementary Material online). A description of the different evolutionary models being tested is provided in Materials and Methods. In brief, we tested four nonmutually exclusive, main evolutionary scenarios: 1) gene duplication was followed by selection (i.e., ω was significantly different from 1) in one or both gene copies (named non-neutral evolution in table 1); 2) gene duplication was followed by asymmetric evolution between the gene copies, so that one copy evolves at a different evolutionary rate (ω), both of the copies evolve at different rates or all three homologs evolve at different rates; 3) gene duplication was followed by significant differences in evolutionary rates along amino acid sites of the sequences and paralogous branches of the tree; and 4) one or both gene copies underwent PS after duplication. These models were applied to triplets of sequences from the four plant species. Each triplet included two sequences from each of the four plant species postdating gene duplication (paralogs) and classified according to their origin (WGD or SSD) and a third, from a related species, predating the duplication event (ortholog) (fig. 1).

Although differences were apparent in the distribution of evolutionary models between duplicates among species, the similarity in the distribution of evolutionary models between WGD and SSD duplicates within each of the species was striking (table 1). For example, in *A. thaliana* the percentage of "non-neutral" tests supporting that at least one of the gene copies evolves under evolutionary rates significantly different from one, either under positive or purifying selection, that is, subjected to functional/selective constraints, when such duplicates come from WGD (81.79% as in table 1) is of the same order as that when these duplicates were originated from SSD (76.27%, table 1). This similarity is observed in all other species and across all the different evolutionary models tested. In contrast, strong differences could be observed among species, with *P. patens* showing the most notable differences (table 1). For instance, the percentage of tests supporting "non-neutral" evolution for one of the gene copies after duplication was highly variable among species (ranging between 16.08% in *P. patens* and 94.81% in *Z. mays*). Similarly, the percentage of significant tests in which both paralogs coming from WGD evolve "non-neutrally" varies between 0.50% (in the case of *P. patens*, table 1) and 78% for *Z. mays*. Similarly, for paralogs originated from SSD, these percentages vary between 0.93% in *P. patens* and 67.20% in *Z. mays* for paralogs originated from SSD (table 1).

The fraction of duplicates in which a model considering neutral evolution in both gene copies was rejected was low, and almost zero in *P. patens* (table 1). However, it is unlikely that the remaining fraction evolves under strict neutral evolution. Duplicated gene(s) evolving neutrally are expected to undergo nonfunctionalization within a few million years after duplication. For instance, assuming two generations per year, it has been previously estimated that nonfunctionalization of a gene copy would take 3.2 My in *A. thaliana*, though this will also be dependent on generation times and population sizes of the species (Ohno 1970; Lynch and Conery 2000). WGD and SSD times of duplication have been predicted to be larger than this estimate (fig. 1). Moreover, all paralogous sequences, with the only exception of six *P. patens* SSD duplicates, were annotated as functional expressed coding genes.

More likely, our results suggest that most duplicated genes have escaped nonfunctionalization and are evolving new or specialized functions. Together, a number of tests support functional specialization, either occurring through asymmetric rates of evolution (affecting one or more branches in the tree) or by significant differences in the substitution rates throughout one or both gene copies (table 1). These tests are compatible with both relaxed purifying selection and PS. Relaxed purifying selection could promote

Downloaded from http://mbe.oxfordjournals.org/ at CSIC on March 1, 2013

Table 1. Summary of Results from Tests of Evolutionary Models on Triplets.

Ingroup Sp.	Non-Neutral Evolution[a] (%)		Asymmetric Evolution (%)			DSR[d] (%)		PS[e] (%)	
	1	2	1[b]	2[b]	Free[c]	1	2	1	2
WGD									
A. thaliana	81.79	48.46	46.54	8.21	42.44	68.33	52.56	29.23	4.62
P. patens	16.08	0.50	70.35	40.20	68.34	79.40	60.80	25.13	5.03
P. trichocarpa	84.51	54.13	44.80	8.83	36.82	58.73	32.60	23.88	2.60
Z. mays	94.65	78.45	24.50	6.62	21.23	33.12	14.53	28.09	1.36
SSD									
A. thaliana	76.27	44.07	48.73	10.17	43.22	66.53	49.58	25	2.97
P. patens	19.58	0.93	72.49	41.34	67.44	77.54	61.69	20.44	3.65
P. trichocarpa	80.92	51.01	47.50	10.42	40.46	60.49	33.02	31.53	3.52
Z. mays	88.73	67.20	24.01	7.17	21.52	37.48	21.23	37.19	3.95

NOTE.—The percentage of LRT tests indicating best fit of the alternative model (P values < 0.05). Numbers in the second row reflect the number of paralogous genes used as foreground branches in LRT tests, e.g., 1 indicates tests accepted for one of the paralogous branches, whereas 2 indicates tests accepted for both paralogous branches. Free indicates the free-ratios branch model (one ratio for each branch). *df*, degrees of freedom.

[a]LRT for "non-neutral" evolution: two-ratios Model 2 ($\omega 0$, $\omega 1 = 1$) vs. two-ratios Model 2 ($\omega 0$, $\omega 1$); $df = 1$.

[b]LRT for asymmetric sequence evolution: one-ratio Model 0 ($\omega 0 = \omega 1 = \omega 2$) vs. two-ratios Model 2 ($\omega 0$, $\omega 1$); $df = 1$.

[c]LRT for asymmetric sequence evolution (free): one-ratio Model 0 ($\omega 0 = \omega 1 = \omega 2$) vs. free-ratios Model 1 ($\omega 0$, $\omega 1$, $\omega 2$); $df = 2$.

[d]LRT for differences in substitution rates (DSR): Model 3 (discrete) vs. clade Model D ($K = 3$); $df = 1$.

[e]LRT for PS: Model A null ($\omega 2 = 1$) vs. Model A ($0 < \omega 0 < 1$); $df = 1$.

sub-functionalization by the balanced distribution of degenerative mutations between duplicates (Duplication, Degeneration, Complementation model). PS is compatible with both neo-functionalization and sub-functionalization through the Escape from Adaptive Conflict model, though further analyses are required to evaluate the relative contribution of these two models. Indeed, a substantial fraction of duplicated genes also supported the corresponding test of PS (table 1). However, better estimates on the role of PS in duplicates specialization would require the analysis of complete gene families including representatives of additional species.

Taken as a whole, these results suggest that higher evolutionary rates observed in SSD duplicates than in WGD duplicates are dependent on the differential intensity of natural selection affecting both gene copies after duplication and is species dependent rather than being the result of a skewed distribution of evolutionary models by mechanism of duplication.

Functional Divergence, Protein Network Connectivity, and Translational Selection after WGD and SSD

The dosage balance hypothesis predicts that preservation of SSD duplicates is selectively disadvantageous because this would generate a stoichiometric unbalance that would interfere with highly constrained cellular systems or genetic networks (Papp et al. 2003; Freeling and Thomas 2006). However, rapid divergence of one gene copy from its ancestral function and its participation in another, less fundamental, cellular system or genetic network could counteract dosage unbalance. Following dosage balance hypothesis, SSD duplications of genes locating to less dense protein interaction network regions and genes expressed at lower levels, are more likely to lead to the preservation of both gene copies through innovation and functional divergence.

We screened duplicated genes formed by either WGD or SSD for evidence of functional divergence. To do so, we measured the differences in the amino acid distances of each paralogous sequence compared with its ancestor (supplementary table S3, Supplementary Material online). We used this measure as a proxy of functional divergence under the assumption that gene copies that are more identical at the sequence level are more likely, on average, to perform more similar functions. Functional divergence of gene copies varied substantially between the different genomes irrespective of the origin of duplication (WGD or SSD) (fig. 3*a*). In three of the four genomes examined, functional divergence was significantly higher between SSD than between WGD duplicates (fig. 3*a*). The higher probability for divergence and functional innovation in SSD duplicates can be due to their lower constraints to evolve new functions. *P. patens* showed a contrasting pattern, maybe reflecting 1) the higher efficiency of selection in removing recessive mutations from haploid genomes and 2) the highly efficient DNA repair mechanisms by homologous recombination of moss, likely reflecting the specific needs of a haploid genome for genome integrity surveillance (Rensing et al. 2008).

We further examined whether SSD duplicates displayed less PPIs, supporting thereby the role of interactions in constraining the evolvability of duplicated genes. We used a library of PPI data for 7,177 *A. thaliana* proteins, obtained by merging results from both experimental and bioinformatic approaches. Assuming that divergence in PPIs of *A. thaliana* orthologs is not biased by the species or the mechanism of duplication, these data were projected on the corresponding orthologs in the other three species under exam (supplementary table S4, Supplementary Material online). Contrary to expectation, differences in the number of PPIs between WGD and SSD proteins were found not to be statistically significant in any of the four species (fig. 3*b*).

Downloaded from http://mbe.oxfordjournals.org/ at CSIC on March 1, 2013

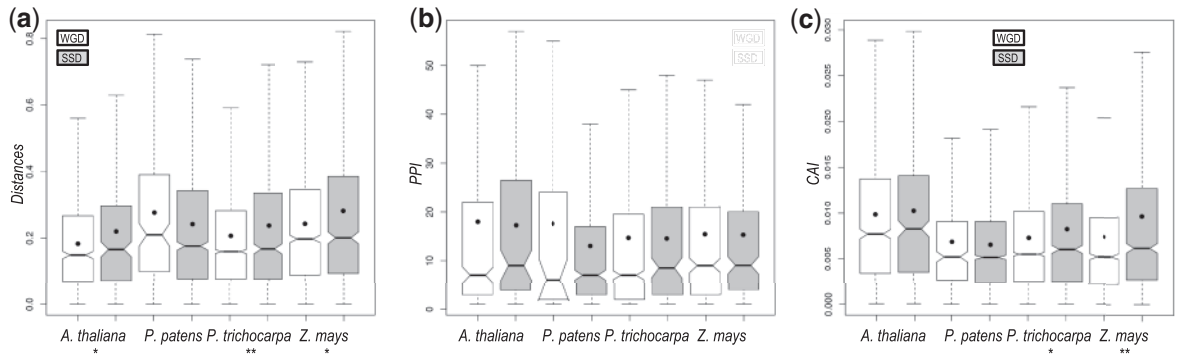

FIG. 3. Notched box-plot representation of functional divergence, number of PPI and differences in CAI between duplicates. Black dots indicate means. All differences were normalized by the magnitude of the values. Differences were tested using a Student's *t* test and the resulting probabilities (*P* values *$<1E-2$, **$<1E-4$) are shown.

To test predictions of the dosage balance hypothesis concerning protein expression, we analyzed differences (normalized increments) in CAI between SSD and WGD gene duplicates (supplementary table S5, Supplementary Material online). CAI measures synonymous codon usage bias of a gene towards the codons used frequently in the most expressed genes of a given species and can be used as a good proxy of translational efficiency and accuracy and, consequently, of the average rates of gene translation under all environmental conditions (Sharp and Li 1987). The average differences in CAI values between duplicated genes were significantly higher in duplicates formed by SSD than in those formed by WGD in *P. trichocarpa* and, specially, *Z. mays* (fig. 3*c*). These differences may contribute to explain the greater divergence levels between duplicates from SSD than those from WGD in these species, as the rate of evolution is strongly correlated with gene expression levels (Drummond et al. 2005, Drummond and Wilke 2008). In the remaining two species, however, the observed difference in the levels of functional divergence between WGD and SSD duplicates would not be influenced by differences in translational rates. Previous results showed higher and broader expression levels in *A. thaliana* duplicates arising from WGD than those originated by SSD as measured from microarray and massively parallel signature sequencing data (Ganko et al. 2007; Yang and Gaut 2011), indicating that divergence is more likely to affect transcriptional regulation.

Synonymous codon usage bias is thought to be dependent on the balance between mutational bias, random genetic drift and natural selection on translational rates (Duret 2002). To examine the relative contribution of the latter, we performed LRT tests of the mutation-selection models (table 2). All of them rejected neutral evolution ($P < 0.05$; unpublished data), providing statistical evidence that synonymous codon usage bias and, consequently, translational rates between paralog sequences, is influenced by natural selection. The proportion of advantageous mutations was quite similar between species and modes of duplication, whereas selection coefficients were slightly variable and within the range of that previously estimated for animal (Yang and Nielsen 2008) and *A. thaliana* genes (dos Reis and Wernisch 2009) (table 2). However, selection on synonymous codon usage is mostly weak and additional features such as differences in population size, recombination rates and variations in GC content should be taken into account to clarify the role of natural selection on species-specific differences in translational efficiency and accuracy between WGD and SSD paralogs.

Table 2. Estimates of Selection Coefficients on Synonymous Codons.

Ingroup Sp.	P+ [a]	|Ns| [b]	Ns+ [c]	Ns− [d]
WGD				
A. thaliana	0.380	0.753	0.386	−0.977
P. patens	0.399	0.656	0.327	−0.873
P. trichocarpa	0.381	0.765	0.382	−1.001
Z. mays	0.368	0.706	0.449	−0.856
SSD				
A. thaliana	0.378	0.751	0.395	−0.967
P. patens	0.399	0.607	0.329	−0.792
P. trichocarpa	0.382	0.737	0.375	−0.962
Z. mays	0.373	0.694	0.431	−0.851

[a]P+, proportion of advantageous mutations.
[b]|Ns|, average selection coefficients of all.
[c]Ns+, advantageous mutations.
[d]Ns−, deleterious mutations.

Differential Patterns of Retention of Functional Categories after SSD and WGD

Understanding the distribution of genes retained in duplicate across the different functional categories in the cell is fundamental from two perspectives: 1) it provides a view of what processes are more prone to evolve novel functions or regulatory activities; and 2) it may allow us to examine the role of the mechanism of duplication in the retention of genes involved in specific functional categories. Previous studies in *A. thaliana* have shown that a substantial number of duplicates were retained in genes involved in fundamental cellular functions: including transcription factors, kinases, certain enzymes and transporters (Blanc and Wolfe 2004a; Maere et al. 2005). However, the propensity of the different cellular functions to retain duplicates and the link between the mode and the fate of duplication remain elusive.

Downloaded from http://mbe.oxfordjournals.org/ at CSIC on March 1, 2013

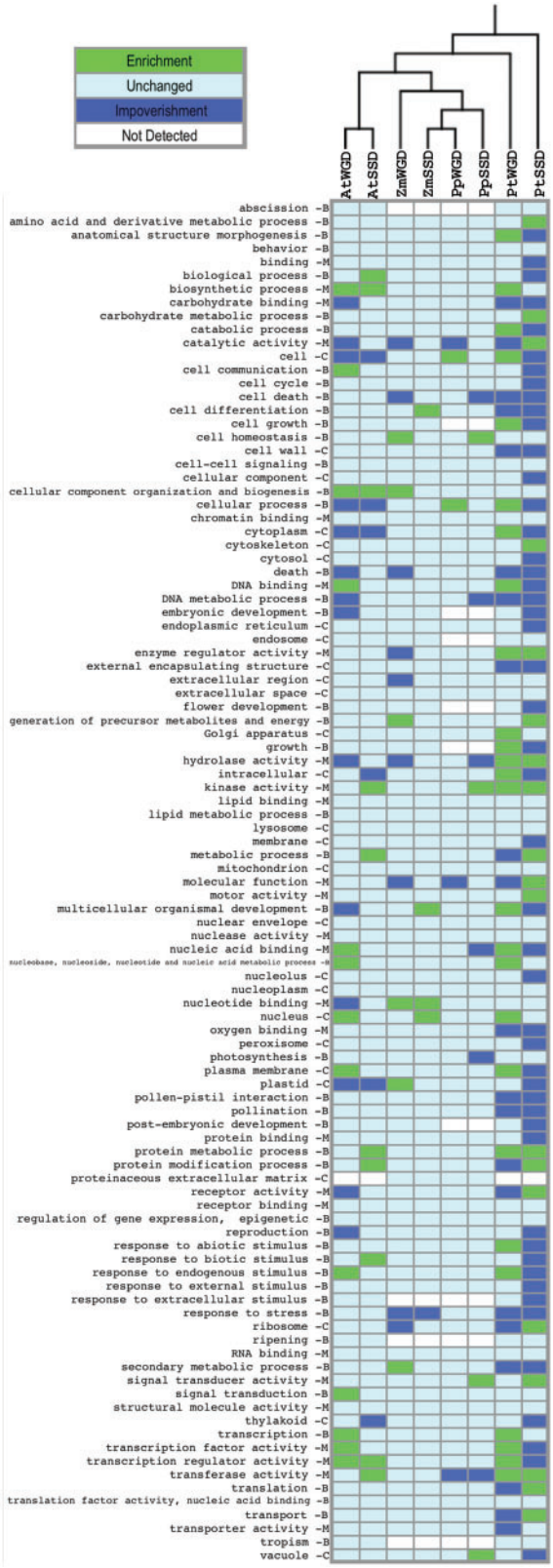

FIG. 4. Clustered color representation of differential retention of plant GO slim terms among duplicates generated through WGD and SSD in four plant genomes. A term was considered as significantly enriched or impoverished in a given data set with respect to the full complement of genes in each genome when Fisher's exact tests returned Bonferroni-corrected P values $< 5 \times 10^{-4}$. B, M, and C indicate GO terms belonging to Biological Process, Molecular Function, and Cellular Component categories, respectively. At, Pp, Pt, and Zm denote *A. thaliana*, *P. patens*, *P. trichocarpa*, and *Z. mays*, respectively. Counts and statistical tests are provided as supplementary information (supplementary table S6, Supplementary Material online). Species were clustered according to their enrichment patterns (represented as a tree on top of the colored matrix). Root is the mean point of the tree and does not represent a species root.

To test the importance of the mode of duplication in determining the dependence between gene functions and its propensity to persist in duplicate, we assigned GO terms to each duplicated gene. Then, we used Fisher's exact tests to identify GO terms that were either enriched or impoverished for duplicated genes originated by WGD or SSD (supplementary table S6, Supplementary Material online). A substantial fraction of duplicated genes was associated with GO terms (ranging from 62.38% to 87.36% for *P. patens* WGD and *A. thaliana* SSD duplicates, respectively). On average, 29% of gene copies shared GO terms. This percentage was slightly higher in WGD (31%) than in SSD paralogs (27%), except for *P. patens* (data not shown), maybe reflecting the stronger degree of functional divergence in gene duplicates generated through SSD. We identified up to 90 out of a total of 103 plant GO categories enriched for genes in duplicate in at least one genome (fig. 4, supplementary table S6, Supplementary Material online). In 19 of these categories, there was no preferential retention of duplicates formed by either WGD or SSD.

Several patterns emerge from the analysis of GOs category enrichments. A few functional categories shared the same retention patterns across species (fig. 4). First, the category of catalytic activities is impoverished for WGD, but not for SSD, duplicates in all four genomes. Conversely, this category was not constrained to retain SSD duplicates and was enriched in *P. trichocarpa* (fig. 4). Second, duplicates belonging to the category kinases were preferentially retained in all genomes except in *Z. mays*, and preferentially so after SSD than WGD. Moreover, *P. trichocarpa* showed enrichment of this category for duplicates originated from both WGD and SSD (fig. 4). Third, molecular function is highly constrained to retain duplicates generated by WGD in all genomes excepting *A. thaliana*, while showing preferential retention among *P. trichocarpa* SSD duplicates (fig. 4). These patterns suggest an important role for the mechanism of duplication in the retention of genes belonging to these functional categories.

In agreement with previous observations (Blanc and Wolfe 2004a; Maere et al. 2005), the categories of transcription (regulation) and signal transduction were specifically enriched for duplicates generated by WGD in *A. thaliana*. We had no evidence, however, pointing to a preferential retention of duplicates originated by either duplication mechanism in these categories in any of the other genomes, with the exception of WGD-formed duplicates in the category transcription (regulation), which were preferentially retained in *P. trichocarpa* (fig. 4).

Indeed, enrichment patterns mainly differ among the examined genomes (fig. 4). In *A. thaliana*, transcription regulator activity, biosynthesis and cellular component organization

Downloaded from http://mbe.oxfordjournals.org/ at CSIC on March 1, 2013

and biogenesis were prone to retain genes after WGD and SSD whereas cytoplasm, cell, cellular processes, and plastid GO categories were highly constrained from retaining WGD or SSD duplicates. In *Z. mays*, nucleotide binding was highly enriched for WGD- and SSD-duplicates retention, while stress response was highly constrained. One category, which was also enriched in *A. thaliana*, was enriched for WGD duplicates in *Z. mays* (cellular component organization and biogenesis). In *P. patens*, genes duplicated through WGD involved in cell and cellular processes were preferentially retained, while four other categories were enriched for SSD duplicates (cell homeostasis, vacuoles, kinase, and signal transducer activities). However, we did not find any category sharing the same pattern of retention with *A. thaliana*. These patterns suggest a prevalent role for the species in the retention of genes belonging to these functional categories.

We further studied the particular distribution of GOs categories in retained duplicates formed by WGD and SSD in *P. trichocarpa* (fig. 4). We found that enriched functional categories were highly related to each other and somewhat complementary between modes of duplication. For example, the enriched categories for WGD duplicates nucleus, nucleic acid and DNA binding, transcription (regulation), nucleotide process, and biosynthesis form a group of functionally related categories related to transcription, while cellular process, (cell) growth, anatomical structure morphogenesis, and multicellular organismal development form another involved in development. In the case of SSD, the enriched categories translation, amino acid and derivative metabolic processes, ribosome and protein metabolism, and modification process would be linked to translation while transport, kinase, receptor and signal transducer activity, and molecular function are involved generically in post-translational regulation. Remarkably, most GO categories enriched for WGD duplicates in *P. trichocarpa* were significantly impoverished with SSD duplicates and vice versa, suggesting a coordinated role of WGD and SSD in *P. trichocarpa* evolutionary diversification, in agreement with the dosage balance hypothesis. *P. trichocarpa* and, to a lesser extent, *A. thaliana* are the two genomes exhibiting the greatest amounts of enrichment, and impoverishment, for duplicate retention, inviting speculation about the higher complexity of evolutionary mechanisms allowing the fixation of gene copies in eudicots.

In summary, the contribution of the species and mechanism of duplication to the retention of duplicates depends on the functional context of the genes. Therefore, the patterns of duplicate retention are not universal but are the result of a complex contribution of the organism lineage, mechanism of duplication and gene function. Taken as a whole, our results indicate that the set of retained duplicates reflect the selective pressures imposed by the ecological requirements of the species rather than being dependent only on the mechanism of duplication.

Conclusion

Here, we have presented a comprehensive analysis of the differential evolutionary dynamics underlying retention of duplicates from WGD and SSD in four independent plant genomes. Duplicates formed by WGD show greater constraints of evolution while SSD paralogs present evidence of increased evolutionary rates and functional divergence after duplication, in concert with the dosage-balance hypothesis. However, examination of the distribution of evolutionary models and functional categories among genes retained in duplicates reveal that the propensity of a gene to persist in duplicate in the genome is not only dependent on the mechanism of duplication but also is heavily influenced by the species, likely linked to the biology and ecological requirements of the plant species. In addition, we provide evidence in support of the differential action of selection on translational rates on both sets of duplicates for the generation of novel functions. Additional work on the role of factors such as genome size, genome compositional bias, chromosomal location, mode of reproduction, population size, and generation time on the patterns of retention and evolution of duplicated genes, that takes into account the results of this study, will shed further light on the evolutionary fates of duplicated genes. Our results highlight the need for studying different species to identify general trends in the evolution of duplicated genes.

Supplementary Material

Supplementary tables S1–S6 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work has been performed making extensive use of PERL scripts and Bioperl and R packages. The authors thank the Bioperl community for continuous support. They are especially grateful to Ken Wolfe, Santiago F Elena, David L Robertson, and Manuel Rodríguez-Concepción for critical reading of the manuscript. This work was supported by a grant from the Spanish Ministerio de Ciencia e Inovación (BFU2009-12022) and a grant of the Research Frontiers Program (10/RFP/GEN2685) from Science Foundation Ireland.

References

Aharoni A, Gaidukov L, Khersonsky O, McQ Gould S, Roodveldt C, Tawfik DS. 2005. The 'evolvability' of promiscuous protein functions. *Nat Genet.* 37:73–76.

Altschul SF, Madden TL, Schaffer 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.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815.

Ashburner M, Ball CA, Blake JA, et al. (20 co-authors). 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 25:25–29.

Bielawski JP, Yang Z. 2004. A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. *J Mol Evol.* 59:121–132.

Downloaded from http://mbe.oxfordjournals.org/ at CSIC on March 1, 2013

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. 2004a. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell* 16: 1679–1691.

Blanc G, Wolfe KH. 2004b. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* 16:1667–1678.

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

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* 17: 540–552.

Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet.* 9:938–950.

Cui L, Wall PK, Leebens-Mack JH, et al. (13 co-authors). 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Res.* 16:738–749.

De Bodt S, Maere S, Van de Peer Y. 2005. Genome duplication and the origin of angiosperms. *Trends Ecol Evol.* 20:591–597.

dos Reis M, Wernisch L. 2009. Estimating translational selection in eukaryotic genomes. *Mol Biol Evol.* 26:451–461.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A.* 102:14338–14343.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134:341–352.

Duret L. 2002. Evolution of synonymous codon usage in metazoans. *Curr Opin Genet Dev.* 12:640–649.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.

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.

Francino MP. 2005. An adaptive radiation model for the origin of new gene functions. *Nat Genet.* 37:573–577.

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

Ganko EW, Meyers BC, Vision TJ. 2007. Divergence in expression between duplicated genes in Arabidopsis. *Mol Biol Evol.* 24:2298–2309.

Geisler-Lee J, O'Toole N, Ammar R, Provart NJ, Millar AH, Geisler M. 2007. A predicted interactome for Arabidopsis. *Plant Physiol.* 145: 317–329.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol.* 11: 725–736.

Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li W-H. 2003. Role of duplicate genes in genetic robustness against null mutations. *Nature* 421:63–66.

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.

He X, Zhang J. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* 169:1157–1164.

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

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci.* 8: 275–282.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res.* 110:462–467.

Li WH, Gu Z, Wang H, Nekrutenko A. 2001. Evolutionary analyses of the human genome. *Nature* 409:847–849.

Lynch M. 2007. The origins of genome architecture. Sunderland (MA): Sinauer Associates.

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

Lynch M, O'Hely M, Walsh B, Force A. 2001. The probability of preservation of a newly arisen gene duplicate. *Genetics* 159: 1789–1804.

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 U S A.* 102:5454–5459.

Moore RC, Purugganan MD. 2005. The evolutionary dynamics of plant duplicate genes. *Curr Opin Plant Biol.* 8:122–128.

Ohno S. 1970. Evolution by gene duplication. New York: Springer-Verlag.

Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. *Nature* 246:96–98.

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

Proost S, Van Bel M, Sterck L, Billiau K, Van Parys T, Van de Peer Y, Vandepoele K. 2009. PLAZA: a comparative genomics resource to study gene and genome evolution in plants. *Plant Cell* 21: 3718–3731.

Rensing SA, Ick J, Fawcett JA, Lang D, Zimmer A, Van de Peer Y, Reski R. 2007. An ancient genome duplication contributed to the abundance of metabolic genes in the moss *Physcomitrella patens*. *BMC Evol Biol.* 7:130.

Rensing SA, Lang D, Zimmer AD, et al. (70 co-authors). 2008. The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science* 319:64–69.

Sharp PM, Li WH. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15:1281–1295.

Simillion C, Vandepoele K, Van Montagu MC, Zabeau M, Van de Peer Y. 2002. The hidden duplication past of *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A.* 99:13627–13632.

Sterck L, Rombauts S, Jansson S, Sterky F, Rouze P, Van de Peer Y. 2005. EST data suggest that poplar is an ancient polyploid. *New Phytol.* 167:165–170.

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.

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

Xia X. 2007. An improved implementation of codon adaptation index. *Evol Bioinform Online.* 3:53–58.

Yang L, Gaut BS. 2011. Factors that contribute to variation in evolutionary rate among Arabidopsis genes. *Mol Biol Evol.* 28: 2359–2369.

Downloaded from http://mbe.oxfordjournals.org/ at CSIC on March 1, 2013

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol.* 19:908–917.

Yang Z, Nielsen R. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol Evol.* 25:568–579.

Zhang J. 2003. Evolution by gene duplication: an update. *Trends Ecol Evol.* 18:292–298.

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

Zhi-Liang H, Bao J, Reecy J. 2008. CateGOrizer: a web-based program to batch analyze gene ontology classification categories. *Online J Bioinformatics.* 9:108–112.

Downloaded from http://mbe.oxfordjournals.org/ at CSIC on March 1, 2013