The Evolution of Gene Duplicates in Angiosperms and the Impact of Protein–Protein Interactions and the Mechanism of Duplication

Jonas Defoort [1,2,3], Yves Van de Peer [1,2,3,4,*], and Lorenzo Carretero-Paulet [1,2,3,*]

[1]Department of Plant Biotechnology and Bioinformatics, Ghent University, Belgium

[2]VIB Center for Plant Systems Biology, Ghent, Belgium

[3]Bioinformatics Institute Ghent, Ghent University, Belgium

[4]Department of Biochemistry, Genetics and Microbiology, University of Pretoria, South Africa

*Corresponding authors: E-mails: yves.vandepeer@psb.ugent.be; lcpaulet@gmail.com, locar@psb.ugent.be.

Accepted: July 10, 2019

Abstract

Gene duplicates, generated through either whole genome duplication (WGD) or small-scale duplication (SSD), are prominent in angiosperms and are believed to play an important role in adaptation and in generating evolutionary novelty. Previous studies reported contrasting evolutionary and functional dynamics of duplicate genes depending on the mechanism of origin, a behavior that is hypothesized to stem from constraints to maintain the relative dosage balance between the genes concerned and their interaction context. However, the mechanisms ultimately influencing loss and retention of gene duplicates over evolutionary time are not yet fully elucidated. Here, by using a robust classification of gene duplicates in *Arabidopsis thaliana*, *Solanum lycopersicum*, and *Zea mays*, large RNAseq expression compendia and an extensive protein–protein interaction (PPI) network from *Arabidopsis*, we investigated the impact of PPIs on the differential evolutionary and functional fate of WGD and SSD duplicates. In all three species, retained WGD duplicates show stronger constraints to diverge at the sequence and expression level than SSD ones, a pattern that is also observed for shared PPI partners between *Arabidopsis* duplicates. PPIs are preferentially distributed among WGD duplicates and specific functional categories. Furthermore, duplicates with PPIs tend to be under stronger constraints to evolve than their counterparts without PPIs regardless of their mechanism of origin. Our results support dosage balance constraint as a specific property of genes involved in biological interactions, including physical PPIs, and suggest that additional factors may be differently influencing the evolution of genes following duplication, depending on the species, time, and mechanism of origin.

Key words: protein–protein interaction, expression divergence, whole genome duplication, small-scale duplication, duplicate retention, angiosperms.

Introduction

Because of the prominent role attributed to gene duplication in generating evolutionary novelty and adaptation, helping to overcome ecological challenges and contributing to the emergence of relevant agronomic traits, the molecular mechanisms driving the evolutionary and functional fate of genes after duplication have been the object of intense research (De Bodt et al. 2005; Conant and Wolfe 2008; Carretero-Paulet and Fares 2012; Panchy et al. 2016; Soltis and Soltis 2016; Van de Peer et al. 2017). Gene duplicates can be broadly classified into two groups based on the size of the genomic region affected by the duplication. Either they result from whole genome duplications (WGDs), also known as polyploidizations, involving the entire genome and thus affecting all genes in the genome, or they originate form small-scale duplications (SSDs), restricted to small genomic regions and mostly involving one to a few genes. Both WGDs and SSDs are highly prevalent among flowering plants (Van de Peer et al. 2009a, 2017; Vanneste et al. 2014), making them perfect models to study evolution after gene duplication. Although most WGDs are followed by intense fractionation (gene loss) and/or genomic rearrangements, removing much of the duplicated genetic features, successful WGDs can be traced back at the base of main plant lineages (Jiao et al. 2011; Amborella

Genome Project 2013), but see also Ruprecht et al. (2017), while more recent WGDs occurred independently in many lineages (Van de Peer et al. 2009a; Vanneste et al. 2014; Soltis and Soltis 2016). For example, in the widely used plant model species *Arabidopsis thaliana*, four WGD events have been detected throughout its evolution (Blanc et al. 2003; Bowers et al. 2003). The most recent ones, namely α and β events, are specific to the Brassicaceae family of rosid eudicots to which *Arabidopsis* belong, whereas the older ones, designated as γ and ε WGD events, are specific to the eudicot and angiosperm lineages, respectively (Jaillon et al. 2007; Amborella Genome Project 2013). Likewise, the asterid eudicot *Solanum lycopersicum* (tomato), a model fruit crop, shares the γ and ε duplication events with *Arabidopsis* and has undergone a more recent whole genome triplication estimated to have occurred around 64 Ma (Tomato Genome Consortium 2012). Finally, also the monocot *Zea mays* (maize) bears traces of several WGD events, the most recent one dated around 5–12 Ma, after divergence with its close relative *Sorghum bicolor* (Blanc and Wolfe 2004b; Schnable et al. 2009). In turn, SSDs can have different origins, including tandem gene duplication and TE-mediated duplication or retroduplication, the most common one being tandem duplication originating from unequal crossing-over resulting in clusters of linearly arranged genes with no or few intervening gene sequences (Panchy et al. 2016). Together with WGD duplicates, tandem duplicates represent the vast majority of duplicates in plants (Panchy et al. 2016).

Previous studies have reported notable differences in the evolutionary and functional fate of duplicates depending on the mechanism or mode of duplication. For example, genes with certain biological functions (e.g., transcriptional regulation, signal transduction, protein transport, and protein modification) are preferentially retained after WGD, whereas they are rarely retained after SSD, and vice versa (Blanc and Wolfe 2004a; Maere et al. 2005a; Carretero-Paulet and Fares 2012; Rodgers-Melnick et al. 2012; Chen et al. 2013; Jiang et al. 2013; Li et al. 2016; Rody et al. 2017). This patterns seems to be universally true because it has also been observed for fungi and vertebrates (Hakes et al. 2007; Wapinski et al. 2007; Makino and McLysaght 2012). Among the different models proposed to explain such biased pattern of loss and retention of duplicates, only the dosage balance hypothesis is claimed to predict such reciprocity between WGD and SSD duplicates (Freeling and Thomas 2006; Freeling 2009; Birchler and Veitia 2014; Conant et al. 2014). The dosage balance hypothesis states that genomes evolve in such a way that encoded proteins forming part of molecular networks and multiprotein complexes or that involved in multiple steps of biological or regulatory pathways, must remain in optimal balance. It is assumed that WGD duplicates do not upset stoichiometry in the cell because all genes in the genome are duplicated simultaneously. Therefore, WGD duplicates will be preferentially retained, as their loss is expected to lead to a dosage imbalance. Conversely, SSD results in one, or few additional gene copies that are likely to upset dosage balance—at least when part of multiprotein complexes or intricate gene regulatory networks—and result in fitness defects, and thus SSD duplicates are expected to be gradually inactivated and deleted from the genome (Lynch and Conery 2000; Conant and Wolfe 2008; Panchy et al. 2016). However, dosage balance is not indefinitely active, and other forces may be at play to explain longer retention times of duplicates (Conant et al. 2014), including selection on absolute gene dosage if higher expression is selectively beneficial (Hudson et al. 2011; Van de Peer et al. 2017), mutational robustness conferred by genetic redundancy (Gu et al. 2003; Keane et al. 2014), interference in the formation of homomultimeric complexes of paralogs harboring degenerative mutations, that is, dominant negatives (Kaltenegger and Ober 2015), or prolonged opportunity for functional specialization to occur (Lynch and Conery 2000; Conant and Wolfe 2008; Conant et al. 2014; Panchy et al. 2016).

The dosage balance hypothesis predicts that reciprocally retained genes are more constrained to evolve novel or specialized functions in order not to upset the dosage balance. Such a prediction was confirmed among *Arabidopsis* gene families classified as dosage balance sensitive using a modeling approach, which were shown to exhibit stronger sequence divergence (SD) constraints and lower rates of functional and expression divergence (ED) (Tasdighian et al. 2017). In agreement with this, 1) duplicates in *Arabidopsis* and poplar resulting from the relatively recent Brassicaceae- and salicoid-specific WGD events, respectively, display lower divergence in expression than tandem duplicates (Casneuf et al. 2006; Rodgers-Melnick et al. 2012), 2) duplicated genes belonging to functional classes and metabolic pathways that are putatively dosage sensitive based on duplication history exhibited reduced expression variance across species after the shared WGD in the *Glycine* lineage (Coate et al. 2016), and 3) WGD duplicates were found to evolve under stronger purifying selection than contemporary SSD duplicates (Yang and Gaut 2011; Carretero-Paulet and Fares 2012; Rodgers-Melnick et al. 2012). Similar differences between duplicates according to their mechanism of duplication could also be observed in the yeast *Saccharomyces cerevisiae*, with WGD duplicates being functionally less different from one another than SSD duplicates (Hakes et al. 2007; Fares et al. 2013). In contrast, Wang et al. reported that WGD duplicates in *Arabidopsis* and rice show greater divergence in expression than tandem duplicates, although differences in the latter were not found to be significant (Wang et al. 2011).

Some findings referring to the impact of protein–protein interactions (PPIs) on duplicate gene evolution are less readily anticipated by the dosage balance hypothesis. For example, a substantial number of WGD duplicates from *Arabidopsis* have diverged in PPI partners, with conservation

declining steadily with the age of the WGD (Guo et al. 2013). Indeed, only a minor fraction of duplicates from the most recent WGD event in *Arabidopsis* involved in PPIs share the same duplication status. The authors claim that the retention of a majority of duplicated gene pairs is no longer explainable by requirements to maintain dosage balance with their interaction partners. Furthermore, although WGD duplicates from *Arabidopsis* and humans display more protein interactions in PPI networks than SSD ones and singletons, differences are only significant for recent duplicates of genes specific to plants or metazoans, respectively (D'Antonio and Ciccarelli 2011; Alvarez-Ponce and Fares 2012). Interestingly, such relationship between centrality in PPI networks and duplicability is inverted in *Escherichia coli*, yeast, worm, and fly (D'Antonio and Ciccarelli 2011). In order to increase our understanding in how PPIs, as well as the mode of duplication, affect gene retention, and the subsequent evolutionary and functional fate of duplicates following WGD and SSD, we here examined a curated data set of WGD and SSD duplicates in *Arabidopsis*, tomato, and maize, a large RNAseq expression compendium with uniquely mapped reads, and an extensive *Arabidopsis* PPI network. Our results point to a key role for PPIs in contributing to dosage balance sensitivity of genes, ultimately helping to explain the biased loss and retention patterns of WGD versus SSD duplicates.

Materials and Methods

Delineation of Gene Families and Identification of Gene Duplicates

Gene families and gene duplicates were delineated and identified for *Arabidopsis*, tomato, maize, and 34 additional flowering plant species as previously described (Li et al. 2016), on the basis of a newly PLAZA 3.0 instance (Proost et al. 2015). The workflow ascribes genes to gene families while homologous regions within and between genomes were identified using i-ADHoRe 3.0 (Proost et al. 2012), with 5 as the minimum number of genes required to define a homologous genomic region as collinear (anchor_points 5), 30 as the maximum number of genes between gene pairs to be considered tandem duplicates (tandem_gap 30), and the rest of settings as reported (Van Bel et al. 2012). Duplicates were further classified as block or tandem duplicates depending on whether they were located in collinear regions of the genomes or were found in the same genomic region as clusters of tandemly arranged genes within a maximum of 30 genes apart, respectively.

Estimates of Synonymous and Nonsynonymous Substitution Rates

For each pair of duplicated genes, codon sequences were aligned with PRANK (version 100701) using the empirical codon model (Kosiol et al. 2007) (setting `-codon`) to align coding DNA, always skipping insertions (`-F`). Estimates of synonymous (K_s) and nonsynonymous substitution rates (K_n) were obtained using the CODEML program in the PAML package (v4.8) (Yang 2007) under the GY model with stationary codon frequencies empirically estimated by the F3 × 4 model (Goldman and Yang 1994). To avoid suboptimal estimates because of maximum likelihood entrapment in local maxima, each calculation was repeated five times, and estimates resulting in the better likelihood were used. Also, in order to reduce the influence of genetic redundancy and of synonymous substitutions saturation from old duplicates, duplicates with a K_s lower than 0.05 and higher than 5, respectively, were discarded from further study (Vanneste et al. 2013).

RNAseq Compendia and Expression Measures

The *Arabidopsis* RNAseq expression compendium was downloaded from Cornet 3.0 and consists of precompiled expression data sets grouping a total 56 experiments (supplementary table S1, Supplementary Material online) (Van Bel and Coppens 2017). The tomato and the maize RNAseq expression compendia were, in turn, taken from the NCBI's Sequence read archive and comprise 84 and 77 different experiments, respectively (supplementary tables S2 and S3, Supplementary Material online). Experiments included a mixture of stress conditions, tissue samples, and developmental stages. The three expression data sets were analyzed using the following pipeline: Trimmomatic 0.30 (Bolger et al. 2014) was first used to perform quality filtering and adaptor removal of the sequencing reads. The reads were then mapped using GSNAP 2015-06-23 (Wu et al. 2016), only retaining uniquely mapped reads. Gene counting was subsequently done using Htseq-count 0.6.1 (Anders et al. 2015), and the resulting counts further transformed to counts per million using EdgeR 3.12.1 (Robinson et al. 2010). To ensure data quality, low expression filtering was performed by removing genes with a sum expression count over all conditions lower than two times the number of total conditions. In total, 19,318 *Arabidopsis*, 19,495 tomato, and 23,164 maize genes were uniquely mapped. The ED between duplicated genes was defined as the relative number of conditions in which only one of the duplicates is detected (C_1 and C_2), divided by the total number of conditions in which they are detected (C).

$$\mathrm{ED} = \frac{C_1 + C_2}{C}.$$

This measure considers the number of conditions in which the duplicates are expressed and reduces differences due to the combination of different experiments. A measure of 0 means that both duplicates are always expressed in the same conditions. A measure of one means that the duplicates were never detected together.

PPI Data

A compendium of PPIs in *Arabidopsis* was constructed combining the following sources: BioGRID 3.4 (Chatr-Aryamontri et al. 2013), *Arabidopsis* Interactome (Arabidopsis Interactome Mapping Consortium 2011), MIND (Jones et al. 2014), CORNET 3.0 (only experimentally validated interactions) (De Bodt et al. 2012), STRING v9.1 (only category binding) (Franceschini et al. 2013), EVEX (http://evexdb.org/) (Van Landeghem et al. 2013) (only category binding), and a data set resulting from transporter associated with antigen processing experiments assembled from literature (Takahashi et al. 2008; Pauwels et al. 2010; Van Leene et al. 2010; Bassard et al. 2012; Eloy et al. 2012; Antoni et al. 2013; Cromer et al. 2013; Di Rubbo et al. 2013; Heijde et al. 2013; Spinner et al. 2013; Cuellar Perez et al. 2014; Fonseca et al. 2014; Gadeyne et al. 2014; Vercruyssen et al. 2014). After removing redundant and self-interactions, we obtained a set of 52,613 interactions for 10,266 proteins. The interaction divergence (ID) between two *Arabidopsis* duplicates was calculated as one minus the retention rate, which in turn was defined as two times the number of interaction partners shared between two duplicates ($I_{1,\ 2}$) divided by the sum of total interactions in each of the duplicates (I_1, I_2).

$$\mathrm{ID} = 1 - \frac{2I_{1,\ 2}}{I_1 + I_2}.$$

In order to categorize tomato and maize duplicates as establishing PPIs or not, *Arabidopsis* PPIs were transferred onto the corresponding orthologous genes in tomato and maize according to the genome-wide gene family classification of these three species together with 34 additional flowering plant species (Li et al. 2016). If at least one interaction was present in one of the *Arabidopsis* genes, all tomato and maize co-orthologous genes in the corresponding gene family were assigned to the category with PPI.

Functional Enrichment Analysis

Enrichment of Gene Ontology (GO) functional terms was calculated using BINGO 2.44 (Maere et al. 2005b), the *Arabidopsis* gene association file from TAIR (GOC Validating Date: March 31, 2017) and the goslim_plant subset version 1.2 (Gene Ontology Consortium 2015). We used hypergeometric and Fisher's exact tests with a *P* value threshold of 0.05 after Benjamini and Hochberg (BH) correction for multiple testing (Benjamini and Hochberg 1995).

Results

Classification of Gene Duplicates, Expression Data Mapping, and PPIs in *Arabidopsis*, Tomato, and Maize

A total of 5,232, 6,645, and 10,654 pairs of duplicated genes were identified in *Arabidopsis*, tomato, and maize, respectively. Duplicates (i.e., ohnologs or homeologs) located in collinear regions of the genomes were further classified as block duplicates putatively arising from WGD events, whereas duplicates found in a singular genomic region were identified as tandem duplicates, conforming the majority of SSD duplicates (table 1). The duplicates that were marked to be both tandem and block duplicates and the ones that could not be unambiguously assigned to any duplication mode were labeled "unclassified" and discarded from further analysis.

We used an expression data set consisting of a compendium of RNAseq experiments for *Arabidopsis*, tomato, and maize (supplementary tables S1–S3, Supplementary Material online). The reads were uniquely mapped and low expression filtering was applied to ensure data quality. Unlike previous studies, where mostly microarray expression data with a low detection rate of duplicates were used (Casneuf et al. 2006; Ganko et al. 2007; Wang et al. 2011; Rodgers-Melnick et al. 2012; Jiang et al. 2013), RNAseq expression data with unique mappings allowed us to individually detect most of the duplicated genes in a pair. In contrast, ATH1 *Arabidopsis* microarrays lacked probes to detect both genes in 38% of duplicate pairs, likely because of cross-hybridization (supplementary table S4, Supplementary Material online). After unique mapping of the reads, expression values were found for both duplicated genes in 63%, 44%, and 48% of *Arabidopsis*, tomato, and maize pairs, respectively. We observed significantly more block duplicates in which both genes in the pair were represented in terms of expression data (79–84%) than tandem duplicates (27–33%) (hypergeometric tests *P* values: *Arabidopsis* $P = 1.16 \times 10^{-183}$, tomato $P = 2.20 \times 10^{-55}$, and maize $P = 4.72 \times 10^{-84}$) (supplementary fig. S1 and supplementary table S5, Supplementary Material online). Tandem duplication is a continuously on-going process, and

Table 1
Distribution of Tandem and Block Duplicates with and without PPIs in *Arabidopsis*, Tomato, and Maize

	Tandem			Block			Unclassified			
	Total	With PPI	Without PPI	Total	With PPI	Without PPI	Total	With PPI	Without PPI	Total
Arabidopsis	1,130	396	734	1,919	1,308	611	2,183	1,199	984	5,232
Tomato	1,534	350	1,184	1,077	693	384	4,034	1,519	2,515	6,645
Maize	1,692	262	1,430	3,400	1,884	1,516	5,562	1,524	4,038	10,654

very recent duplicates are expected to show little or null SD, likely resulting in the observed higher number of young tandem duplicates without unique expression read mapping.

Finally, we assembled a compendium of *Arabidopsis* PPIs based on small- and large-scale experiments. A total of 2,903 *Arabidopsis* duplicates were found as involved in PPIs. Tomato and maize duplicates were further categorized as involved in PPIs or not by projecting PPI data from *Arabidopsis* duplicates onto their corresponding orthologous genes in these two species, using the genome-wide gene family classification of 37 species of flowering plants (Li et al. 2016). A total of 2,562 and 3,670 pairs of duplicates with PPIs in at least one member of the pair were predicted in tomato and maize, respectively (table 1).

Block Duplicates Evolve Slower than Tandem Duplicates

Previous studies on *Arabidopsis* and poplar duplicates supported that the mechanism of duplication resulted in differential constraints to evolve, with WGD duplicates generally evolving under stronger purifying selection (Yang and Gaut 2011; Carretero-Paulet and Fares 2012; Rodgers-Melnick et al. 2012) or displaying lower divergence in expression than tandem ones (Casneuf et al. 2006; Rodgers-Melnick et al. 2012). In order to test, and eventually confirm these observations with our three-species data set, we calculated measures of divergence at the level of sequence (SD) and expression (ED) for each of the duplicate pairs in all three species. The rates of nonsynonymous substitutions (K_n), resulting in amino acid changes, were used as estimates of SD between duplicates and also, indirectly, as a proxy for functional divergence (Fares et al. 2013). In turn, ED was calculated as the relative number of conditions in which only one of the duplicates is detected.

First, we examined the relationship among K_s, SD, and ED, as well as the putative influence of the mechanism of duplication, by performing pairwise Pearson and Spearman rank correlation tests among these variables for duplicates in all three species partitioned by mechanism of duplication. It had been previously suggested that correlation of ED with K_s only occurred among younger duplicates (Wang et al. 2011). To account for this, we generated a second subset of younger duplicates restricted to those with estimates of $K_s < 1$. In all three species and for both modes of duplication and subsets of duplicates, we found a strongly significant positive correlation between K_s and SD both through Pearson and Spearman rank tests (fig. 1 and supplementary table S6, Supplementary Material online). With respect to ED, a positive correlation with K_s was only found among block duplicates, although r were generally pretty low (supplementary table S6, Supplementary Material online). In turn, among tandem duplicates, only a marginally significant positive correlation was found between K_s and ED in *Arabidopsis*, being nonsignificant in tomato, or even marginally negative in the

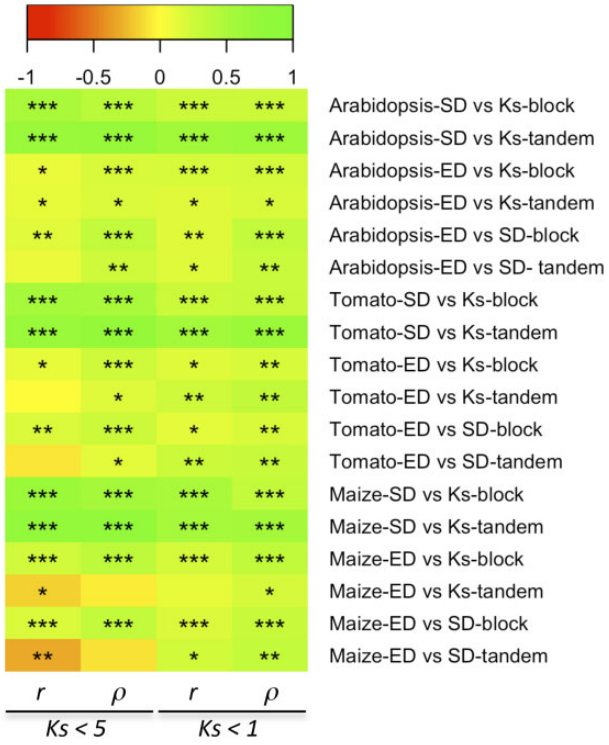

Fig. 1.—Heat map of pairwise correlation analysis among K_s, SD (K_n), and ED in *Arabidopsis*, tomato, and maize duplicates partitioned by mechanism of duplication (block vs. tandem). Pearson's (r) and Spearman's rank (ρ) correlation coefficients resulting from comparing subsets of duplicates with $K_s < 5$ or $K_s < 1$ are colored according to the legend, and the significance level (***, $<10x - 10$; **, $<10x - 5$; *, <0.05) of the associated P values are shown.

case of maize (fig. 1 and supplementary table S6, Supplementary Material online). Similar results were obtained between SD (K_n) and ED, with only block duplicates displaying a significant positive correlation, whereas tandem ones showed no significant correlation, or a negative one as in the case of maize (fig. 1 and supplementary table S6, Supplementary Material online). Interestingly, although Spearman's rank tests generally resulted in better correlation coefficients and P values, no significant negative correlation was found for any subset of duplicates and comparison performed. Similarly, we found no significant negative correlation in any comparison when we restricted our analysis to duplicates showing $K_s < 1$. Taken as a whole, these results seem to indicate that the occurrence of species-specific outlier duplicates with high K_s values would be altering the linear relationship between SD and ED found for younger duplicates and support previous observations about the heterogeneous relationship between SD and nucleotide substitutions (Wang et al. 2011).

We further studied the impact of the mechanism of duplication on the evolution of SD and ED over time, using K_s as a proxy of evolutionary time. As synonymous substitutions do

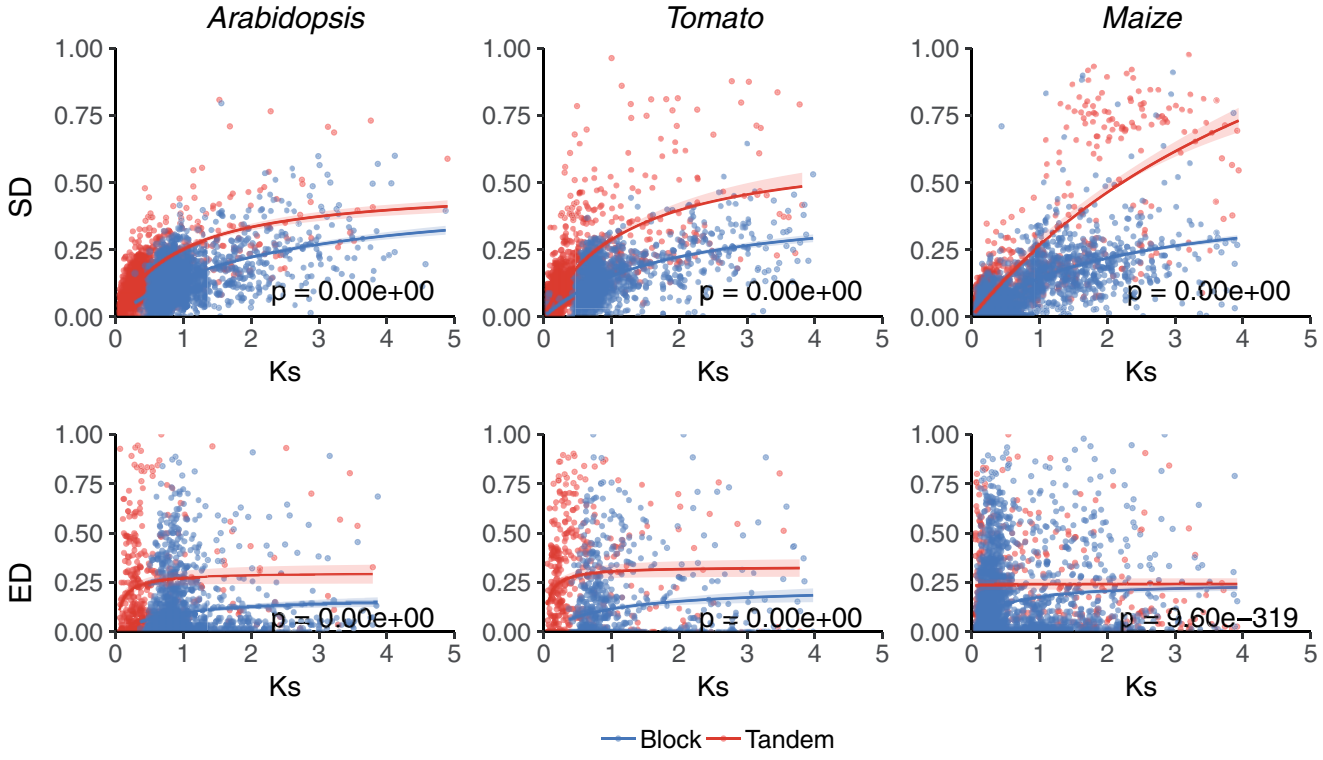

Fig. 2.—Evolution of sequence (SD) and expression (ED) divergence of tandem and block duplicates in *Arabidopsis*, tomato, and maize. SD (upper panels) and ED (lower panels) plotted as a function of K_s. For every species, Michaelis–Menten-type saturation curves were fit to SD or ED values of tandem and block duplicates independently. Ninety five percent confidence regions are indicated as colored areas around the corresponding curves. The *P* values on the plots result from *F*-tests for fitting two Michaelis–Menten-type curves independently for tandem and block versus one curve to the combined data set of all duplicates (data not shown).

not result in amino acid changes, they are not supposed to impact the function and/or structure of the resulting encoded protein, consequently accumulating throughout evolution in a (nearly) neutral manner. Because of the low coefficients obtained in the correlation analysis, especially between ED and K_s or K_n, together with the weak, negative, or nonlinear relationship observed in some species and subsets of duplicates, linear regression did not seem the most appropriate function to model the evolution of SD and ED of duplicates. Furthermore, saturation at K_s values >1 caused by the gradual accumulation of multiple substitutions at the same site over time is not fully corrected for by current evolutionary models and may lead to spurious results (Vanneste et al. 2013). Therefore, we opted for Michaelis–Menten type saturation curves, which had already been proven successful (Tasdighian et al. 2017) in modeling K_s saturation for old(er) duplicates. Assuming functional redundancy at the time of duplication (i.e., ED and SD should be 0), we model the putative impact of the mechanism of duplication over evolution by plotting our estimates of ED and SD between duplicates as a function of K_s, and fitting independent Michaelis–Menten type saturation curves to tandem and block duplicates. Significance of the differences of the variances between subsets of duplicates was assessed through *F*-tests for testing the hypothesis of fitting two curves independently versus a simpler nested model in which one curve was fitted to the combined data set. As shown in figure 2, ED and SD of *Arabidopsis*, tomato, and maize block duplicates putatively arising from WGD events were consistently found to diverge significantly slower over time than tandem duplicates.

We next explored whether the mechanism of duplication could also be constraining the evolution of divergent PPI partners using measures of ID between *Arabidopsis* duplicated genes. We restricted our analysis to *Arabidopsis*, for which we had assembled a compendium of experimentally determined PPI data. ID was calculated as 1 minus the retention rate, defined as the number of interaction partners shared between two duplicates divided by the sum of unique interaction partners of both duplicates. In order to reduce the noise due to the high rate of false negatives (i.e., not all proteins have experimental PPI data), ID was only calculated for duplicates in which one of the duplicates has at least four PPIs and the other duplicate at least one PPI. Seven hundred and eighty eight pairs were found to be above this cutoff. There are more block duplicates (23%) with more than half of the interaction partners conserved, compared with only 6% for tandem duplicates (Fisher's exact test: $P = 1.2 \times 10^{-8}$). We also found more tandem duplicates without any shared

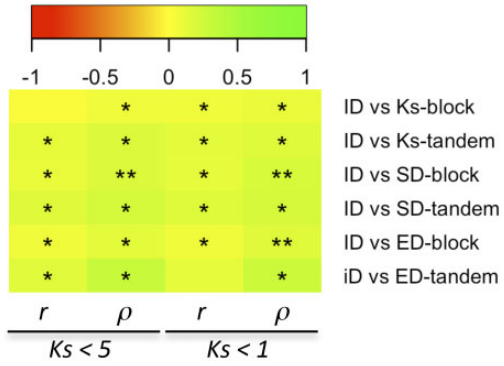

Fig. 3.—Heat map of correlation analysis between K_s, SD (K_n), and ED versus ID in *Arabidopsis*, partitioned by mechanism of duplication (block vs. tandem). Pearson's (r) and Spearman's rank (ρ) correlation coefficients resulting from comparing the subsets of duplicates with $K_s < 5$ or $K_s < 1$ are colored according to the legend, and the significance levels (***, $<10x - 10$; **, $<10x - 5$; *, <0.05) of the associated P values are shown.

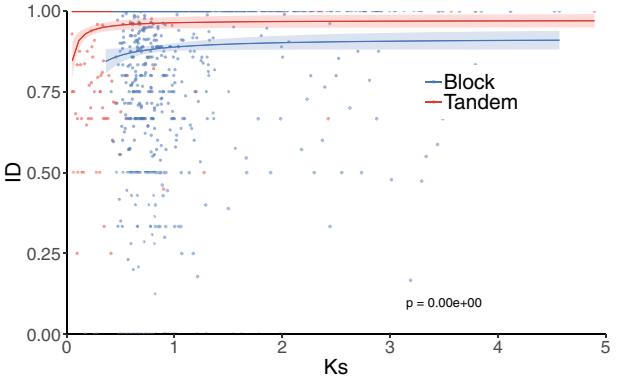

Fig. 4.—Evolution of ID of *Arabidopsis* tandem and block duplicates. ID for pairs of *Arabidopsis* duplicates plotted as a function of K_s. Michaelis–Menten-type saturation curves were fit to ID values of tandem and block duplicates independently. Ninety five percent confidence regions are indicated as colored areas around the corresponding curves. The P values on the plots result from F-tests for fitting two Michaelis–Menten-type curves independently for tandem and block versus one curve to the combined data set of all duplicates (data not shown).

interaction partners (48%) than block duplicates (30%) (Fisher's exact test: $P = 2.3 \times 10^{-2}$). Correlations between ID and K_s or ID and K_n were positive and generally significant, although only marginally, especially in the latter. The linear relationship between ID and K_s or by K_n is weak, as reflected by the low coefficients obtained (fig. 3 and supplementary table S7, Supplementary Material online). A marginally significant positive correlation was also found between ID and ED (fig. 3 and supplementary table S7, Supplementary Material online). Finally, we plotted ID as a function of evolutionary time and fitted independent Michaelis–Menten curves to block and tandem duplicates. The former appeared to be significantly more constrained to gain or loss different PPI

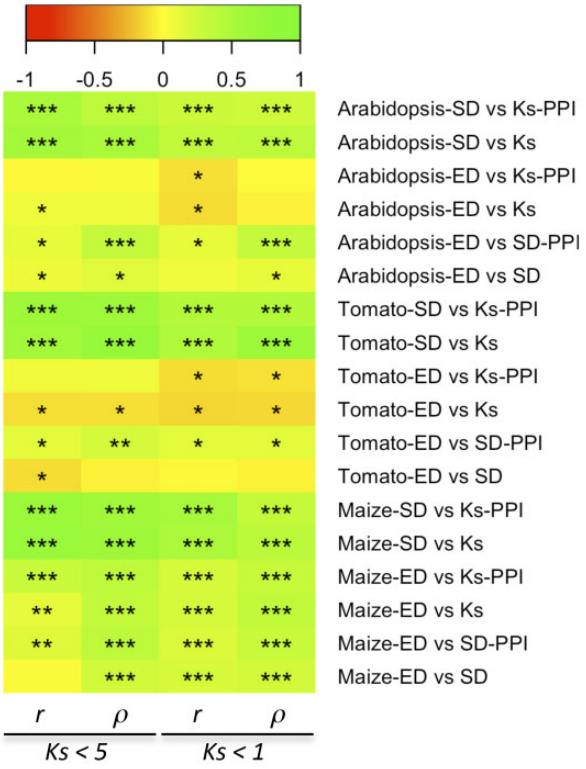

Fig. 5.—Heat map of pairwise correlation analysis between K_s, SD (K_n), and ED in *Arabidopsis*, tomato, and maize duplicates partitioned by PPI category (without PPI vs. with PPI). Pearson's (r) and Spearman's rank (ρ) correlation coefficients resulting from comparing the subsets of duplicates with $K_s < 5$ or $K_s < 1$ are colored according to the legend, and the significance levels (***, $<10x - 10$; **, $<10x - 5$; *, <0.05) of the associated P values are shown.

partners than the later, an effect that persists over time (fig. 4). Our analyses were replicated using different cutoffs to assign a pair to the category with PPI (from at least one up to 14 interaction partners in one of the duplicates), always resulting in significant differences between tandem and block duplicates (supplementary fig. S2, Supplementary Material online).

Duplicates with PPIs Are More Constrained to Evolve Divergent Functions

To investigate the putative impact of PPIs on the functional and evolutionary divergence of duplicates, we first examined pairwise correlations among K_s, SD, or ED between duplicates from all three species, partitioned by the PPI category to which the duplicate belongs to (i.e., duplicates without PPI vs. duplicates with PPI), and for two subsets of duplicates (with $K_s < 5$ and $K_s < 1$). Both Pearson and Spearman rank tests showed a strongly significant positive correlation between K_s and K_n in all three species for both PPI categories and subsets of duplicates (fig. 5 and supplementary table S8, Supplementary

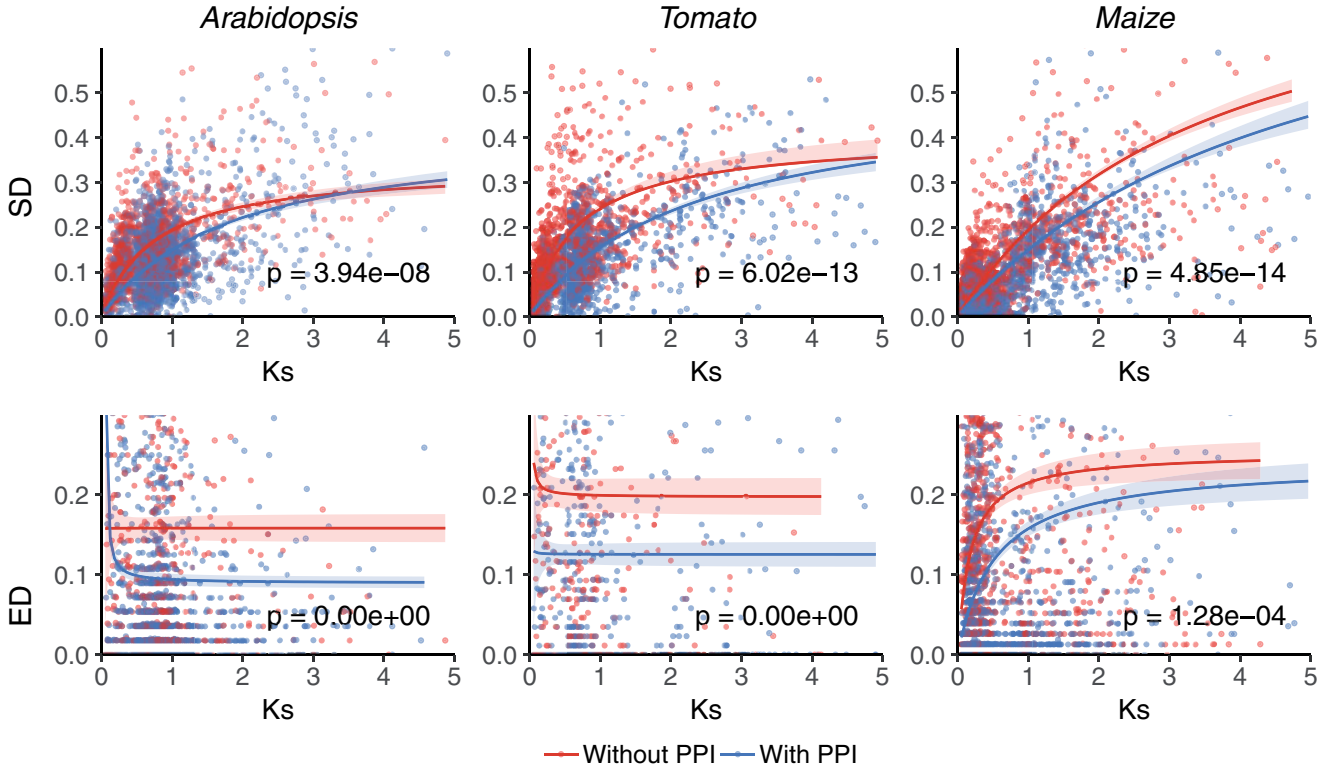

Fig. 6.—Evolution of sequence (SD) and expression (ED) divergence of duplicates with and without PPI in *Arabidopsis*, tomato, and maize. SD (upper panels) and ED (lower panels) plotted as a function of K_s. For every species, Michaelis–Menten-type saturation curves were fit to SD or ED of duplicates with and without PPIs independently. Ninety five percent confidence regions are indicated as colored areas around the corresponding curves. The *P* values on the plots result from *F*-tests for fitting two Michaelis–Menten-type curves independently for duplicates with or without PPIs versus one curve to the combined data set of all duplicates (data not shown). In order to improve the interpretability of the results, the *y* axes were truncated at 0.6 and 0.3 for SD and ED, respectively.

Material online). In turn, correlation between K_s and ED was generally low, nonsignificant, or even negative such as in the case of tomato duplicates without PPIs (fig. 5 and supplementary table S8, Supplementary Material online). When we restricted our analysis to the subset of duplicates with $K_s < 1$, negative correlations between K_s and ED could also be detected among tomato duplicates with PPIs, as well as for both *Arabidopsis* duplicates with and without PPIs. In turn, K_n between duplicates with PPIs showed significant positive correlation with ED in all three species, especially in Spearman rank tests and for duplicates with $K_s < 1$. Among duplicates without PPIs, correlation between SD and ED was found to be not significant, or only marginally positive or negative in *Arabidopsis* and tomato, respectively (fig. 5 and supplementary table S8, Supplementary Material online).

Next, we further examined the putative influence of establishing PPIs in the evolution of ED and SD over time, by plotting our estimates of ED and SD between duplicates with and without PPIs as a function of K_s, and fitting independent Michaelis–Menten type saturation curves to each subset of duplicates. As can be observed in figure 6, ED and SD evolve significantly slower in duplicates with PPIs than in duplicates without PPIs in all three species, suggesting the occurrence of PPIs constraints the evolution of duplicates at the expression pattern and sequence level. This constraint generally seems to persist over long evolutionary times, although this may be obscured in the plots due to the low number of duplicates in the upper K_s region. The constraint on duplicates evolution imposed by PPIs appears to be dependent on the actual number of PPI partners, as reflected their significant negative correlations with SD (Pearson correlation tests: tandem $r = -0.096$, $P = 3.4 \times 10^{-3}$; block $r = -0.18$, $P = 9.7 \times 10^{-16}$) and ED (Pearson correlation tests: tandem $r = -0.19$, $P = 3.4 \times 10^{-4}$; block $r = -0.16$, $P = 3.7 \times 10^{-10}$) of *Arabidopsis* duplicates.

Block and Tandem Duplicates with PPIs Evolve Slower than Their Counterparts without PPIs

With the aim of exploring the interplay between the occurrence of PPIs and the mechanism of duplication in the evolution of ED and SD between duplicates, we plotted estimates of ED and SD for pairs of *Arabidopsis*, tomato, and maize duplicated genes over K_s by separately partitioning tandem

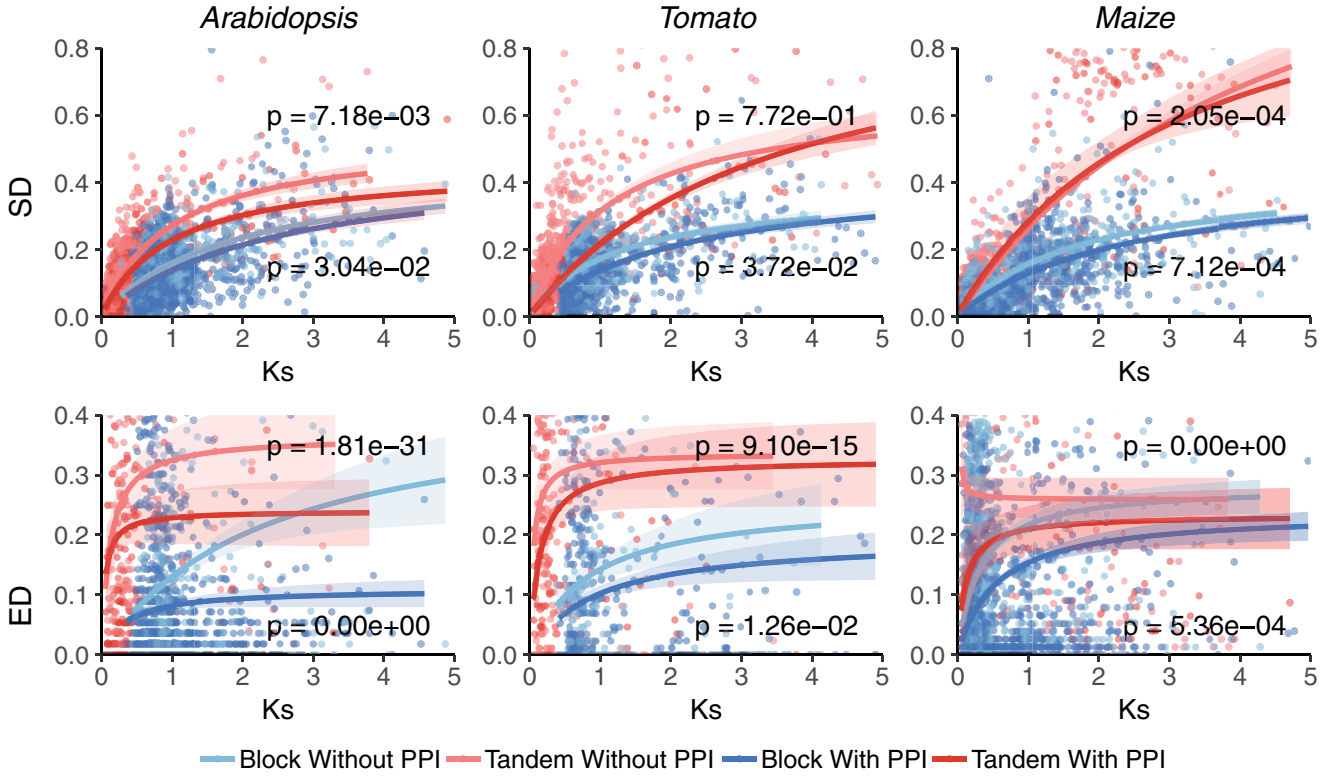

Fig. 7.—Evolution of sequence (SD) and expression (ED) divergence of tandem and block duplicates with and without PPIs in *Arabidopsis*, tomato, and maize. SD (upper panels) and ED (lower panels) plotted as a function of K_s. For every species, Michaelis–Menten-type saturation curves were fit to SD and ED values of tandem or block duplicates with and without PPIs independently. Ninety five percent confidence regions are indicated as colored areas around the corresponding curves. The *P* values on the plots result from *F*-tests for fitting two Michaelis–Menten-type curves independently for duplicates with PPIs and without PPIs within each duplication mode versus one curve to the combined data set of all duplicates of each duplication mode (data not shown). In order to improve the interpretability of the results, the *y* axes were truncated at 0.8 and 0.4 for SD and ED, respectively.

and block duplicates with and without PPIs, and fitted independent Michaelis–Menten type saturation curves to each subset of duplicates. We then performed *F*-tests for fitting two Michaelis–Menten-type curves independently for either tandem or block duplicates with and without PPIs versus one curve to the combined data set of duplicates of each kind (fig. 7). Eleven out of 12 *F*-tests resulted in significant differences in ED or SD between both tandem and block duplicates with and without PPIs (fig. 7). The general picture that emerges is that of duplicates without PPIs displaying faster rates of ED and SD evolution than their counterparts with PPIs.

We next investigated the distribution of PPIs between modes of duplication (table 1). In all three species, PPIs were found to be strongly overrepresented among block duplicate genes (Fisher's exact tests with BH correction: *Arabidopsis* $P = 3.07 \times 10^{-37}$, tomato $P = 2.13 \times 10^{-49}$, and maize $P = 5.13 \times 10^{-116}$), whereas underrepresented among tandem ones (Fisher's exact test with BH correction: *Arabidopsis* $P = 5.25 \times 10^{-11}$, tomato $P = 6.68 \times 10^{-8}$, and maize $P = 2.53 \times 10^{-48}$) (table 1). However, the average number of PPI partners of *Arabidopsis* tandem (6.094) and block duplicates (6.300) did not show significant differences (*t*-test: $P = 0.541$), which allows to discard the possibility that the differences observed above could be due to differences in the average number of PPI partners between duplication modes.

Finally, we examined whether PPIs could be also influencing the expected reciprocal pattern of enrichment in GO molecular functions between modes of duplication in *Arabidopsis* (Blanc and Wolfe 2004a; Maere et al. 2005a; Carretero-Paulet and Fares 2012; Rodgers-Melnick et al. 2012; Chen et al. 2013; Jiang et al. 2013; Li et al. 2016; Rody et al. 2017). Block duplicates with PPI were enriched for GO terms associated with binding (protein, nucleic acid, DNA, and RNA), kinase activity (catalytic, transferase), signal transduction/receptor activity, most of which were found as showing no changes or being significantly underrepresented among tandem duplicates with and without PPIs, respectively (fig. 8). This pattern of enrichment contrasted with that of block duplicates without PPIs, where only catalytic activity was similarly overrepresented, together with hydrolase activity, which also popped up as strongly enriched. In turn, tandem

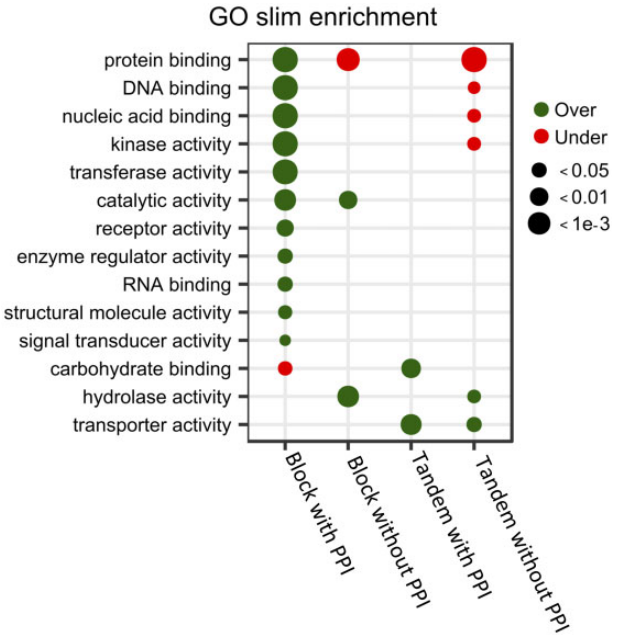

Fig. 8.—Functional enrichment analysis of block and tandem duplicates with and without PPI. Enrichment analysis of GO molecular functions belonging to the plant GO slim category for *Arabidopsis* block and tandem duplicates with and without PPI. Only experimentally validated GO annotations were considered. GO terms significantly under- and over-represented (P value < 0.05 hypergeometric test with BH correction) are plotted.

duplicates were enriched for transporter activity, with carbohydrate binding and hydrolase activity also found as specifically enriched among those with or without PPIs, respectively.

Discussion

Here, we have studied the impact of the mechanism of duplication and of PPIs on the evolutionary and functional fate of gene duplicates in three angiosperm plants with different histories of SSD and WGD. By using uniquely mapped RNAseq compendia, we were able to detect the majority of the duplicates in a more robust and reliable way compared with previous studies using microarray data (Casneuf et al. 2006; Ganko et al. 2007; Wang et al. 2011; Rodgers-Melnick et al. 2012; Jiang et al. 2013), although there is still some room for improvement to detect young tandem duplicates in the lower K_s regions. Furthermore, we assembled a massive compendium of PPI data in *Arabidopsis* and tried to overcome the lack of experimental PPI data in other plant species by projecting our *Arabidopsis* PPI network onto the corresponding orthologs in tomato and maize, with the purpose of categorizing them as establishing PPIs or not. Although orthologous proteins in different species may have evolved divergent functions, including the gain and loss of specific interaction partners, we followed the conservative approach of transferring PPI data between gene families, instead of individual genes. Although this methodology is not perfect and it is likely to result in a high degree of noise, this is not expected to affect SSD or WGD duplicates differently, introducing a bias in our observations.

Our results support contrasting evolutionary dynamics of functional and evolutionary divergence between block and tandem duplicates in all three species, which are likely reflecting their differential contribution to evolutionary innovation and adaptation. Block duplicates consistently diverge slower in terms of SD and ED, indicating stronger purifying selection to evolve novel or divergent protein functions, expression domains or PPI partners, respectively, that may upset dosage balance with other partners of the affected networks. These differences are likely related to the different mutational mechanisms of each mode of duplication; although WGD duplicates entire genes including cis-regulatory regions, SSD often results in incomplete duplication of the gene owing to the random nature of DNA breakage and recombination (Casneuf et al. 2006; Zou et al. 2009). Furthermore, low or null correlations generally observed between ED and nucleotide substitution rates at the level of coding sequences are likely related to the fact that changes in gene expression patterns also rely on changes in promoter or UTR regions (Wang et al. 2011). Similarly, ID showed stronger constraints to evolve among *Arabidopsis* block than tandem duplicates. This pattern did not seem to originate from differences in the average number of PPIs between modes of duplication, as these were not found to be significant as previously noted in *Arabidopsis* (Carretero-Paulet and Fares 2012) and yeast (Hakes et al. 2007).

Although the slower evolution of block duplicates is anticipated by the dosage balance hypothesis, it also raises the question of the biological and evolutionary significance of WGD or polyploidy. The paucity of successful paleopolyploidy events in extant species suggests that polyploidy is usually an evolutionary "dead end" (Van de Peer et al. 2009b; Mayrose et al. 2011; Van de Peer et al. 2017). However, at specific times in evolution, organisms that underwent and survived WGDs might have had some adaptive advantage over their diploid progenitors, eventually contributing to 1) evolutionary diversification and increase in biological complexity (Van de Peer et al. 2009b; Soltis and Soltis 2016, 2009; Van de Peer et al. 2017), as supported by the polyploidy events observed at the base of main plant lineages (Jiao et al. 2011; Amborella Genome Project 2013), but see also Ruprecht et al. (2017) and 2) successful adaptations under periods of extreme environmental stress and/or fluctuations, as suggested by the wave of lineage-specific WGD events observed in angiosperms around the time of the Cretaceous-Paleogene (K-Pg) extinction event (Fawcett et al. 2009; Van de Peer et al. 2009a, 2017; Vanneste et al. 2014). It has been argued that dosage balance selection against functional specialization of block duplicates might be limiting the role of polyploidy on

promoting evolutionary change (Tasdighian et al. 2017). However, dosage balance constraints are expected to fade away or change over time (Conant et al. 2014), and thus should be viewed as the primary force driving the retention of duplicates shortly after duplication. Block duplicates retained over longer times may provide with prolonged opportunity for neutral subfunctionalization via the Duplication–Degeneration–Complementation model to occur (Force et al. 1999; Conant and Wolfe 2008; Fares et al. 2013). Subfunctionalization also paves the way for subsequent adaptive evolution under positive selection of novel functions (neofunctionalization) or improvement of ancestral secondary functions (subfunctionalization via the Escape from Adaptive Conflict) (He and Zhang 2005; Conant and Wolfe 2008; Des Marais and Rausher 2008; Panchy et al. 2016). Furthermore, the probabilities of rewiring duplicated networks formed by multiple connected proteins into entire novel complex metabolic, regulatory, or developmental pathways increase if all genes involved duplicate together by means of WGD and evolve synchronously novel or specialized subfunctions, such as interactions partners or expression domains. This way, WGD duplicates originally retained neutrally through requirements to maintain dosage balance, can contribute to the complex adaptive changes at the genomic level and the phenotypic plasticity required in the face of events of evolutionary radiation or ecological challenge.

Tandem duplicates are more likely to upset dosage balance, in special when connected with other proteins. Their retention in the short term will depend on the cost associated with the maintenance of additional gene copies. The faster divergence rates observed for tandem duplicates in all three species may thus reflect the rapid acquisition of novel or specialized functions in order to compensate this cost; otherwise, they are expected to be lost by means of nonfunctionalization or pseudogenization (Lynch and Conery 2000). This, together with across-species differences observed in correlation patterns between ED and K_s or K_n for young and old tandem duplicates might suggest their involvement in rapid adaptations to local environmental stimuli, which is in turn supported by species-specific enrichments commonly observed for tandem duplicates in functional categories related to response to stress or secondary metabolism (Hanada et al. 2008; Denoeud et al. 2014; Panchy et al. 2016). Long-term retention of specific duplicates may also result from selection on the absolute dosage of certain gene products, that is, the higher concentration of an enzyme may result in the higher metabolic flux in the cell of the corresponding biochemical pathway (Bekaert et al. 2011; Hudson et al. 2011). This selection is also expected to operate differently on block and tandem duplicates. In pathways where increases in the absolute dosage of a single enzyme have no effect on the resulting metabolic flux, WGDs can provide such a flux increase by duplicating all its components at once (Bekaert et al. 2011). In contrast, enzymes that are working independently or that provide a bottleneck in the pathway could take advantage of a SSD (e.g., hexose transport in yeast) (Sugino and Innan 2006; Arakaki et al. 2011).

Functional and evolutionary divergence of *Arabidopsis*, tomato, and maize duplicates also appeared to be constrained by the involvement of the encoded protein in PPIs, as revealed by the significant slower rates of evolutionary change in terms of SD and ED of duplicates with PPIs. These constraints are dependent on the actual number of PPI partners, as reflected by the low, although significant, negative correlations with SD and ED in both *Arabidopsis* block and tandem duplicates, that is, the higher the number of PPI partners, the higher the constraint for duplicates to diverge. Regions of the protein involved in PPI interactions, that is, PPI interfaces, are conserved through negative purifying selection, which is expected to limit amino acid changes (Lovell and Robertson 2010). Therefore, a given protein involved in multiple PPI interactions is expected to show a reduced number of sequence regions available for evolutionary change to occur without disrupting PPI interfaces, thus resulting in the observed increased selective constraint to diverge. These observations are in agreement with duplicates involved in physical protein–protein, or other molecular or genetic, interactions evolving under stronger purifying selection, because functional divergence of a connected protein is more likely to disrupt the stoichiometry of the affected biological network (Freeling and Thomas 2006; Freeling 2009; Birchler and Veitia 2014; Conant et al. 2014). Furthermore, the fraction of block duplicates with PPIs is significantly larger than that of tandem duplicates, which may be reflecting the fact that the chance of upsetting dosage balance if lost increases for connected WGD duplicates.

Our results also supported PPIs as imposing stronger selective constraints independently of the duplication mode, that is, both block and tandem duplicates with PPIs show slower rates of ED and SD evolution than their counterparts without PPIs. Our functional enrichment analysis further revealed GO molecular functions commonly reported in the literature as associated with dosage sensitive functional classes, that is, transcriptional regulation, development, and signaling (Blanc and Wolfe 2004a; Maere et al. 2005a; Carretero-Paulet and Fares 2012; Rodgers-Melnick et al. 2012; Chen et al. 2013; Jiang et al. 2013; Li et al. 2016; Rody et al. 2017), are specifically enriched among *Arabidopsis* block duplicates with PPIs, with the reciprocal pattern being true for tandem duplicates without PPIs. Interestingly, hydrolase enzymatic activity appeared as enriched in both groups of duplicates without PPIs. Therefore, the reciprocal retention pattern predicted by the dosage balance hypothesis (Freeling and Thomas 2006; Freeling 2009; Birchler and Veitia 2014; Conant et al. 2014) can be, at least partially, explained by the enrichment in PPIs of genes involved in biological functions commonly classified as dosage balance sensitive, rather than by the mechanism of duplication itself. However, it must be noted that the generally low correlation coefficients obtained in our analysis, particularly for ED or ID versus K_s or K_n, are suggesting that other

factors, apart from the mechanism of duplication and PPIs, are affecting the functional divergence of duplicates. These additional factors likely include other biological interactions, apart from physical PPIs, in which the gene, or its product, is involved. Assuming constraints of duplicates to functionally diverge throughout evolution are solely based on dosage balance sensitivity, it is tempting to speculate that subsets of duplicates not involved in any interaction or network, that is, functioning in solitary, if any, will evolve under similar selection regimes. However, the current analysis suggests additional species-specific mechanisms not necessarily influencing dosage balance sensitivity may be at play and highlights the complexity of the mechanisms underlying functional divergence of duplicates throughout evolution (Carretero-Paulet and Fares 2012).

In summary, our results support dosage balance constraints of duplicates to functionally diverge as specific properties of genes, rather than associated with specific biological functions, and resulting from their overall involvement in different kinds of biological interactions and networks. Of these, we have shown the prominent role played by PPIs in explaining differential dosage balance sensitivity and subsequent duplicate retention and contribution to evolutionary innovation and adaptation between modes of duplication. Current progresses on systems biology approaches integrating high-throughput-omics data, together with the development of evolutionary simulation computational frameworks, will help to unravel the contribution of relative dosage balance sensitivity to explain gene evolution after duplication with respect to other models proposed, including absolute dosage balance, functional specialization through neo- or sub-functionalization, mutation robustness, or paralog interference.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported by funding from the European Union Seventh Framework Program (FP7/2007-2013) under European Research Council Advanced Grant Agreement 322739 – DOUBLEUP to Y.V.d.P. We thank Christina Toft for valuable comments and helpful discussions on the manuscript.

Literature Cited

Alvarez-Ponce D, Fares MA. 2012. Evolutionary rate and duplicability in the *Arabidopsis thaliana* protein–protein interaction network. Genome Biol Evol. 4(12):1263–1274.

Amborella Genome Project. 2013. The *Amborella* genome and the evolution of flowering plants. Science 342(6165):1241089.

Anders S, Pyl PT, Huber W. 2015. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics 31(2):166–169.

Antoni R, et al. 2013. PYRABACTIN RESISTANCE1-LIKE8 plays an important role for the regulation of abscisic acid signaling in root. Plant Physiol. 161(2):931–941.

Arabidopsis Interactome Mapping Consortium. 2011. Evidence for network evolution in an *Arabidopsis* interactome map. Science 333:601–607.

Arakaki M, et al. 2011. Contemporaneous and recent radiations of the world's major succulent plant lineages. Proc Natl Acad Sci U S A. 108(20):8379–8384.

Bassard JE, et al. 2012. Protein–protein and protein–membrane associations in the lignin pathway. Plant Cell 24(11):4465–4482.

Bekaert M, Edger PP, Pires JC, Conant GC. 2011. Two-phase resolution of polyploidy in the *Arabidopsis* metabolic network gives rise to relative and absolute dosage constraints. Plant Cell 23(5):1719–1728.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B 57:289–300.

Birchler JA, Veitia RA. 2014. The Gene Balance Hypothesis: dosage effects in plants. Methods Mol Biol. 1112:25–32.

Blanc G, Hokamp K, Wolfe KH. 2003. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. Genome Res. 13(2):137–144.

Blanc G, Wolfe KH. 2004a. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. Plant Cell 16(7):1679–1691.

Blanc G, Wolfe KH. 2004b. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell 16(7):1667–1678.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120.

Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422(6930):433–438.

Carretero-Paulet L, Fares MA. 2012. Evolutionary dynamics and functional specialization of plant paralogs formed by whole and small-scale genome duplications. Mol Biol Evol. 29(11):3541–3551.

Casneuf T, De Bodt S, Raes J, Maere S, Van de Peer Y. 2006. Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant *Arabidopsis thaliana*. Genome Biol. 7(2):R13.

Chatr-Aryamontri A, et al. 2013. The BioGRID interaction database: 2013 update. Nucleic Acids Res. 41(Database issue):D816–D823.

Chen E, et al. 2013. The dynamics of functional classes of plant genes in rediploidized ancient polyploids. BMC Bioinformatics 14(Suppl 15):S19.

Coate JE, Song MJ, Bombarely A, Doyle JJ. 2016. Expression-level support for gene dosage sensitivity in three Glycine subgenus Glycine polyploids and their diploid progenitors. New Phytol. 212(4):1083–1093.

Conant GC, Birchler JA, Pires JC. 2014. Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. Curr Opin Plant Biol. 19:91–98.

Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet. 9(12):938–950.

Cromer L, et al. 2013. Centromeric cohesion is protected twice at meiosis, by SHUGOSHINs at anaphase I and by PATRONUS at interkinesis. Curr Biol. 23(21):2090–2099.

Cuellar Perez A, et al. 2014. The non-JAZ TIFY protein TIFY8 from *Arabidopsis thaliana* is a transcriptional repressor. PLoS One 9:e84891.

D'Antonio M, Ciccarelli FD. 2011. Modification of gene duplicability during the evolution of protein interaction network. PLoS Comput Biol. 7:e1002029.

De Bodt S, Hollunder J, Nelissen H, Meulemeester N, Inzé D. 2012. CORNET 2.0: integrating plant coexpression, protein–protein interactions, regulatory interactions, gene associations and functional annotations. New Phytol. 195(3):707–720.
De Bodt S, Maere S, Van de Peer Y. 2005. Genome duplication and the origin of angiosperms. Trends Ecol Evol (Amst). 20(11):591–597.
Denoeud F, et al. 2014. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science 345:1181–1184.
Des Marais DL, Rausher MD. 2008. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. Nature 454(7205):762–765.
Di Rubbo S, et al. 2013. The clathrin adaptor complex AP-2 mediates endocytosis of brassinosteroid insensitive1 in *Arabidopsis*. Plant Cell 25(8):2986–2997.
Eloy NB, et al. 2012. SAMBA, a plant-specific anaphase-promoting complex/cyclosome regulator is involved in early development and A-type cyclin stabilization. Proc Natl Acad Sci U S A. 109(34):13853–13858.
Fares MA, Keane OM, Toft C, Carretero-Paulet L, Jones GW. 2013. The roles of whole-genome and small-scale duplications in the functional specialization of *Saccharomyces cerevisiae* genes. PLoS Genet. 9(1):e1003176.
Fawcett JA, Maere S, Van de Peer Y. 2009. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. Proc Natl Acad Sci U S A. 106(14):5737–5742.
Fonseca S, et al. 2014. bHLH003, bHLH013 and bHLH017 are new targets of JAZ repressors negatively regulating JA responses. PLoS One 9(1):e86182.
Force A, et al. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151(4):1531–1545.
Franceschini A, et al. 2013. STRING v9.1: protein–protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 41(Database issue):D808–D815.
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(7):805–814.
Gadeyne A, et al. 2014. The TPLATE adaptor complex drives clathrin-mediated endocytosis in plants. Cell 156(4):691–704.
Ganko EW, Meyers BC, Vision TJ. 2007. Divergence in expression between duplicated genes in *Arabidopsis*. Mol Biol Evol. 24(10):2298–2309.
Gene Ontology Consortium. 2015. Gene Ontology Consortium: going forward. Nucleic Acids Res. 43:D1049–D1056.
Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 11(5):725–736.
Gu Z, et al. 2003. Role of duplicate genes in genetic robustness against null mutations. Nature 421(6918):63–66.
Guo H, Lee T-H, Wang X, Paterson AH. 2013. Function relaxation followed by diversifying selection after whole-genome duplication in flowering plants. Plant Physiol. 162(2):769–778.
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(10):R209.
Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu SH. 2008. Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiol. 148(2):993–1003.
He X, Zhang J. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics 169(2):1157–1164.
Heijde M, et al. 2013. Constitutively active UVR8 photoreceptor variant in *Arabidopsis*. Proc Natl Acad Sci U S A. 110(50):20326–20331.
Hudson CM, Puckett EE, Bekaert M, Pires JC, Conant GC. 2011. Selection for higher gene copy number after different types of plant gene duplications. Genome Biol Evol. 3:1369–1380.
Jaillon O, et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467.
Jiang WK, Liu YL, Xia EH, Gao LZ. 2013. Prevalent role of gene features in determining evolutionary fates of whole-genome duplication duplicated genes in flowering plants. Plant Physiol. 161(4):1844–1861.
Jiao Y, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. Nature 473(7345):97–100.
Jones AM, et al. 2014. Border control—a membrane-linked interactome of *Arabidopsis*. Science 344(6185):711–716.
Kaltenegger E, Ober D. 2015. Paralogue interference affects the dynamics after gene duplication. Trends Plant Sci. 20(12):814–821.
Keane OM, Toft C, Carretero-Paulet L, Jones GW, Fares MA. 2014. Preservation of genetic and regulatory robustness in ancient gene duplicates of *Saccharomyces cerevisiae*. Genome Res. 24(11):1830.
Kosiol C, Holmes I, Goldman N. 2007. An empirical codon model for protein sequence evolution. Mol Biol Evol. 24(7):1464–1479.
Li Z, et al. 2016. Gene duplicability of core genes is highly consistent across all angiosperms. Plant Cell 28(2):326–344.
Lovell SC, Robertson DL. 2010. An integrated view of molecular coevolution in protein–protein interactions. Mol Biol Evol. 27(11):2567–2575.
Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science 290(5494):1151–1155.
Maere S, Heymans K, Kuiper M. 2005. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics 21(16):3448–3449.
Maere S, et al. 2005. Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci U S A. 102(15):5454–5459.
Makino T, McLysaght A. 2012. Positionally biased gene loss after whole genome duplication: evidence from human, yeast, and plant. Genome Res. 22(12):2427–2435.
Mayrose I, et al. 2011. Recently formed polyploid plants diversify at lower rates. Science 333(6047):1257.
Panchy N, Lehti-Shiu M, Shiu SH. 2016. Evolution of gene duplication in plants. Plant Physiol. 171(4):2294–2316.
Pauwels L, et al. 2010. NINJA connects the co-repressor TOPLESS to jasmonate signalling. Nature 464(7289):788–791.
Proost S, et al. 2012. i-ADHoRe 3.0—fast and sensitive detection of genomic homology in extremely large data sets. Nucleic Acids Res. 40(2):e11.
Proost S, et al. 2015. PLAZA 3.0: an access point for plant comparative genomics. Nucleic Acids Res. 43(Database issue):D974–D981.
Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1):139–140.
Rodgers-Melnick E, et al. 2012. Contrasting patterns of evolution following whole genome versus tandem duplication events in *Populus*. Genome Res. 22(1):95–105.
Rody HV, Baute GJ, Rieseberg LH, Oliveira LO. 2017. Both mechanism and age of duplications contribute to biased gene retention patterns in plants. BMC Genomics. 18(1):46.
Ruprecht C, et al. 2017. Revisiting ancestral polyploidy in plants. Sci Adv. 3(7):e1603195.
Schnable PS, et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. Science 326(5956):1112–1115.
Soltis PS, Soltis DE. 2009. The role of hybridization in plant speciation. Annu Rev Plant Biol. 60:561–588.
Soltis PS, Soltis DE. 2016. Ancient WGD events as drivers of key innovations in angiosperms. Curr Opin Plant Biol. 30:159–165.
Spinner L, et al. 2013. A protein phosphatase 2A complex spatially controls plant cell division. Nat Commun. 4:1863.

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(12):642–644.
Takahashi N, et al. 2008. The DNA replication checkpoint aids survival of plants deficient in the novel replisome factor ETG1. EMBO J. 27(13):1840–1851.
Tasdighian S, et al. 2017. Reciprocally retained genes in the angiosperm lineage show the hallmarks of dosage balance sensitivity. Plant Cell 29(11):2766–2785.
Tomato Genome Consortium. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485:635–641.
Van Bel M, Coppens F. 2017. Exploring plant co-expression and gene–gene interactions with CORNET 3.0. Methods Mol Biol. 1533:201–212.
Van Bel M, et al. 2012. Dissecting plant genomes with the PLAZA comparative genomics platform. Plant Physiol. 158(2):590–600.
Van de Peer Y, Fawcett JA, Proost S, Sterck L, Vandepoele K. 2009. The flowering world: a tale of duplications. Trends Plant Sci. 14(12):680–688.
Van de Peer Y, Maere S, Meyer A. 2009. The evolutionary significance of ancient genome duplications. Nat Rev Genet. 10(10):725–732.
Van de Peer Y, Mizrachi E, Marchal K. 2017. The evolutionary significance of polyploidy. Nat Rev Genet. 18(7):411–424.
Van Landeghem S, et al. 2013. Large-scale event extraction from literature with multi-level gene normalization. PLoS One 8(4):e55814.
Van Leene J, et al. 2010. Targeted interactomics reveals a complex core cell cycle machinery in *Arabidopsis thaliana*. Mol Syst Biol. 6:397.
Vanneste K, Baele G, Maere S, Van de Peer Y. 2014. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. Genome Res. 24(8):1334–1347.
Vanneste K, Van de Peer Y, Maere S. 2013. Inference of genome duplications from age distributions revisited. Mol Biol Evol. 30(1):177–190.
Vercruyssen L, et al. 2014. ANGUSTIFOLIA3 binds to SWI/SNF chromatin remodeling complexes to regulate transcription during *Arabidopsis* leaf development. Plant Cell 26(1):210–229.
Wang Y, et al. 2011. Modes of gene duplication contribute differently to genetic novelty and redundancy, but show parallels across divergent angiosperms. PLoS One 6(12):e28150.
Wapinski I, Pfeffer A, Friedman N, Regev A. 2007. Natural history and evolutionary principles of gene duplication in fungi. Nature 449(7158):54–61.
Wu TD, Reeder J, Lawrence M, Becker G, Brauer MJ. 2016. GMAP and GSNAP for genomic sequence alignment: enhancements to speed, accuracy, and functionality. Methods Mol Biol. 1418:283–334.
Yang L, Gaut BS. 2011. Factors that contribute to variation in evolutionary rate among *Arabidopsis* genes. Mol Biol Evol. 28(8): 2359–2369.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24(8):1586–1591.
Zou C, Lehti-Shiu MD, Thomashow M, Shiu SH. 2009. Evolution of stress-regulated gene expression in duplicate genes of *Arabidopsis thaliana*. PLoS Genet. 5(7):e1000581.

Associate editor: Brian Golding