Evolution of Gene Duplication in Plants

Nicholas Panchy, Melissa Lehti-Shiu, and Shin-Han Shiu*

Genetics Program (N.P., S.-H.S.) and Department of Plant Biology (M.L.-S., S.-H.S.), Michigan State University, East Lansing, Michigan 48824

ORCID IDs: 0000-0002-1551-3517 (N.P.); 0000-0003-1985-2687 (M.L.-S.); 0000-0001-6470-235X (S.-H.S.).

Ancient duplication events and a high rate of retention of extant pairs of duplicate genes have contributed to an abundance of duplicate genes in plant genomes. These duplicates have contributed to the evolution of novel functions, such as the production of floral structures, induction of disease resistance, and adaptation to stress. Additionally, recent whole-genome duplications that have occurred in the lineages of several domesticated crop species, including wheat (*Triticum aestivum*), cotton (*Gossypium hirsutum*), and soybean (*Glycine max*), have contributed to important agronomic traits, such as grain quality, fruit shape, and flowering time. Therefore, understanding the mechanisms and impacts of gene duplication will be important to future studies of plants in general and of agronomically important crops in particular. In this review, we survey the current knowledge about gene duplication, including gene duplication mechanisms, the potential fates of duplicate genes, models explaining duplicate gene retention, the properties that distinguish duplicate from singleton genes, and the evolutionary impact of gene duplication.

Distinct from other eukaryotic genomes, plant genomes tend to evolve at higher rates, leading to higher genome diversity (Kejnovsky et al., 2009; Murat et al., 2012). For example, differences in genome size between closely related plant species are much larger than between other closely related eukaryotes. Among dicotyledonous species that diverged approximately 150 million years ago (MYA), genome size ranges from merely 63 Mb in the carnivorous *Genlisea margaretae* (Greilhuber et al., 2006) to approximately 150 Gb in the canopy plant *Paris japonica* (Pellicer et al., 2010). This 2,000-fold difference in genome size among dicots is in stark contrast to that observed among the mammalian species that also radiated approximately 150 MYA (Warren et al., 2008), where genome size ranges from approximately 1.6 Gb in Carriker’s round-eared bat (Smith et al., 2013) to approximately 8 Gb in the tetraploid red viscacha rat (Gallardo et al., 1999).

Plant genomes also have an abundance of duplicate genes. Whole-genome duplication (WGD) has occurred multiple times over the past 200 million years of angiosperm evolution (Lyons et al., 2008; Soltis et al., 2009, 2014; Lee et al., 2013; Renny-Byfield and Wendel, 2014), and genomic sequencing continues to reveal new events (Velasco et al., 2010; D’Hont et al., 2012; Wang et al., 2012; Lu et al., 2013; Myburg et al., 2014; Wang et al., 2014b). In contrast, the most recent WGD event occurred approximately 450 MYA in the lineage leading to humans (Panopoulou et al., 2003; Dehal and Boore, 2005) and approximately 200 MYA in the budding yeast lineage (Wolfe and Shields, 1997; Kellis et al., 2004). Strikingly, many plant species also comprise mixed populations of diploid and polyploid individuals, illustrating the prevalence of polyploidy in plants (Husband et al., 2013). For example, 2.4% of *Lythrum salicaria* populations have both diploid and polyploid individuals (Kubatova et al., 2008), and this percentage is even higher (greater than 60%) for *Chamerion angustifolium* (Sabara et al., 2013) and *Actinidia chinensis* (Li et al., 2010).

WGD, or polyploidization, is an extreme mechanism of gene duplication that leads to a sudden increase in
both genome size and the entire gene set. However, it is not the only mechanism that gives rise to duplicated genes. In general, gene duplication generates two gene copies; this theoretically allows one or both to evolve under reduced selective constraint and, on some occasions, to acquire novel gene functions that contribute to adaptation. There is little question that duplicate genes have contributed to novel traits over the course of plant evolution (Van de Peer et al., 2009b). Through comparative analyses of an ever-increasing number of plant genome sequences and functional genomic data sets, we now have an unprecedented understanding of how genes are duplicated, how duplicated genes evolve new functions, and the impact of gene duplication on genome evolution (Conant and Wolfe, 2008; Freeling et al., 2015; Soltis et al., 2015).

Gene duplication is but one type of genomic change that can lead to evolutionary novelties. Novel functions can arise from the co-option of existing genes (True and Carroll, 2002), new genes can arise de novo from intergenic space (Tautz and Domazet-Lošo, 2011; Schlötterer, 2015), and new transcriptional regulatory sites can come into existence that alter gene expression (Wray et al., 2003). In addition, although in this review we focus only on genes, the duplication of other genomic features, including regulatory regions (Nourmohammad and Lässig, 2011), transposable elements (TEs; Lisch, 2013), and repeat elements (Sharopova, 2008), has been reported to influence gene expression and function. Nonetheless, gene duplication remains of specific interest both because of the abundance of plant gene duplicates and their potential to contribute to plant novelties. The goal of this review is to provide an overview of our current state of knowledge about plant gene duplication and its significance. We first focus on the prevalence of gene duplication in plants and the mechanisms that contribute to gene duplication. We then discuss the fate of duplicate genes and the factors that influence whether a duplicate is retained or not. Finally, we consider the influence of duplicate genes on the evolution of plant species and agronomically important traits.

PREVALENCE AND MECHANISMS OF DUPLICATION

Predominance of Duplicate Genes in Plant Genomes

In the green lineage, gene numbers range from 8,166 in the unicellular green alga Ostroococcus tauri (Derelle et al., 2006) to approximately 95,000 in bread wheat (Triticum aestivum; Brenchley et al., 2012). What proportion of genes within each species have shared common ancestry due to duplication (Fitch, 1970)? Although a paralog is well defined conceptually, the criteria (e.g. the threshold sequence similarity) and data sets used to identify paralogs vary among studies, and thus direct comparisons are difficult. For example, between 16% (barley [Hordeum vulgare]) and 49% (rice [Oryza sativa]) of plant genes were defined as paralogous based on transcript data (Blanc and Wolfe, 2004a); however, genome sequence data have yielded paralog frequencies as high as approximately 75% in soybean (Glycine max; Schmutz et al., 2010). In Arabidopsis (Arabidopsis thaliana), the estimate of duplicate gene content ranges from 47% (Blanc and Wolfe, 2004a) to 63% (Ambrosino et al., 2016) due to differences in gene models, methodology, and parameters (i.e. similarity cutoffs).

To obtain a comparable estimate of duplicate gene number across plant genomes, we applied a common methodology and similarity threshold to identify duplicate genes in 41 sequenced land plant genomes (Fig. 1). On average 64.5% of plant genes are paralogous, ranging from 45.5% in the bryophyte Physcomitrella patens to 84.4% in apple (Malus domestica). Given that ancient and/or fast-evolving paralogs are not easily detected due to sequence divergence, these percentages are likely underestimates. Total genic content is correlated significantly with both paralog content ($r^2 = 0.46$, $P < 7e-6$) and the presence of a reported polyploidization event ($r^2 = 0.35$, $P < 5e-4$), demonstrating the large contribution of duplication, particularly WGD, to differences in gene content among plant species.

Another way to illustrate the preponderance of plant duplicates is to look at the number of paralogs within gene families (Dayhoff, 1976). In a survey of eight diverse plant species, the percentage of genes belonging to gene families ranges from 40% in the green alga Chlamydomonas reinhardtii to 95% in the lycophyte Selaginella moellendorffii, with most species having in excess of 65% familial genes (Guo, 2013). Although the proportion of familial genes in plant genomes is high, there can be dramatic differences in the size of gene families across species due to lineage-specific expansions (Lespinet et al., 2002). For example, one of the largest families in plants is the protein kinase superfamily, which has 426 members in the unicellular green alga C. reinhardtii and 2,532 in Eucalyptus grandis (Lehti-Shiu and Shiu, 2012). Another example illustrating the variation in plant gene family size is the large difference in the number of transcription factors, which can differ more than 10-fold among plant species (Jin et al., 2014).

At the other extreme, some genes have few or no paralogs. For example, there is only one Arabidopsis gene encoding DNA gyrase A, despite the fact that repeated rounds of WGD would have generated gyrase A duplicates in the past. Thus, not all gene families are created equal. What contributes to these large differences in gene family size? Integrated analysis of gene family and functional annotation data led to the finding that plant genes involved in transcriptional regulation, signal transduction, and stress response tend to have paralogs (Blanc and Wolfe, 2004b; Maere et al., 2005; Shiu et al., 2005; Hanada et al., 2008) but those involved in essential functions, such as genome repair, genome duplication, and organelles, tend not to (Li et al., 2016). The correlation between duplication and function also appears to be influenced by how duplicates are made (Hanada et al., 2008). For example, transcription factors...
have higher than average retention after WGD (Maere et al., 2005) but not after local duplication (Hanada et al., 2008). Therefore, to understand how duplication impacts gene content, it is necessary to know how genes are duplicated.

**Mechanisms of Gene Duplication**

Duplication is a form of mutation in which a genomic region is replicated and, in some cases, is inserted into a physically separate location. Multiple mechanisms contribute to gene duplication (Fig. 2). Considering the impact on gene content, the most dramatic form of gene duplication involves duplication of an entire chromosome or the whole genome (Fig. 2A). Ancient WGD events have taken place in the common ancestors of seed plants (approximately 340 MYA) and of angiosperms (approximately 170 MYA; Jiao et al., 2011). Subsequently, three rounds of WGD events (referred to as α, β, and γ) took place in the Arabidopsis lineage (Blanc et al., 2003; Bowers et al., 2003). In some cases, WGD events involve genome triplication, as is the case for *Brassica rapa* (Lysak et al., 2005), the wild radish...
Raphanus raphanistrum (Moghe et al., 2014), and bread wheat (Salse et al., 2008). Even higher levels of polyploidization also have been observed (e.g. octoploid cultivated strawberry [Fragaria × ananassa]; Byrne and Jelenkovic, 1976). Paleopolyploidy events are thought to be rare, as only a handful of recognizable events occurred over the past 200 MYA (Van de Peer et al., 2009a). Nonetheless, the frequency of paleopolyploidization is much higher in plants than in any other eukaryotic lineage (Otto and Whitton, 2000). In addition, although the survival rate of nascent polyploids remains largely unclear (Soltis et al., 2015), diploid and polyploid cytotypes frequently coexist (Ramsey and Schemske, 1998). In Arabidopsis, for example, approximately 60% of genes have at least one paralog in a corresponding syntenic block derived from one of three WGD events (Bowers et al., 2003).

The contribution of WGD to existing duplicates is likely much higher than estimated based on analyses of syntenic duplicates, because some WGD syntenic blocks, particularly those arising from older events, are no longer recognizable due to genome rearrangements, insertions, and deletions. Thus, some of the duplicates that cannot currently be ascribed to WGD may actually be WGD duplicates. Alternatively, they could be derived from subgenomic duplication events, which include tandem duplication (Zhang, 2003), duplication mediated by TEs (Jiang et al., 2004), segmental duplication (Bailey et al., 2002), and retroduplication (Drouin and Dover, 1990; Brosius, 1991). Tandem (or local) duplication (Fig. 2B) results from unequal crossing-over events and leads to a cluster of two to many paralogous sequences with no or few intervening gene sequences (Zhang, 2003). The number of tandem duplicates in plants varies widely, from 451 (4.6% of gene content) in Craspedia variabilis to 16,602 (26.1%) in apple (Yu et al., 2015). In Arabidopsis, the proportion of tandem

![Mechanisms of gene duplication](https://example.com/mechanisms.png)

**Figure 2.** Mechanisms of gene duplication. A, WGD, or duplication of genes via an increase in ploidy. B, Tandem duplication, or duplication of a gene via unequal crossing over between similar alleles. C, Transposon-mediated duplication, or duplication of a gene associated with a TE via replicative transposition. D, Segmental duplication, or duplication of genes via replicative transposition of LINE elements in the human genome. The extent and causative elements of segmental duplication are ill defined in plants. E, Retroduplication, or duplication of a gene via reverse transcription of processed mRNA.
duplicates is close to the average of approximately 9% observed for 39 plant species. One challenge in estimating the contribution of tandem duplication to duplicate gene content is that, in the case of recent duplication events, paralogous copies may be misannotated as a single gene. For example, there are two SEC10 paralogous genes involved in exocytotic vesicle fusion, but these genes make up only one locus in the assembled Arabidopsis genome (Vukašinović et al., 2014). Thus, misassembly contributes to an underestimate of tandem genes, and this issue is likely significant as most plant genomes are of draft quality.

In contrast to tandem duplication, which takes place locally, other subgenomic duplication mechanisms form dispersed duplicates. These mechanisms likely involve repetitive sequences and/or replicative transposition by TEs (Alleman and Freeling, 1986; Kapitonov and Jurka, 2007). In the case of TE-mediated gene duplication (Fig. 2C) in plants, gene capture by Mutator-like elements (MULEs) is the most prominent example (Bennetzen, 2005). There are approximately 3,000 rice Pack-MULEs that collectively contain approximately 1,000 fragmented or whole-gene sequences (Jiang et al., 2004). There are a similar number of Helitrons (approximately 2,800) in maize (Zea mays; Du et al., 2009) but only 46 Pack-MULEs in Arabidopsis (Jiang et al., 2011). This difference potentially reflects the historical difference in TE activity. Interestingly, Pack-MULEs preferentially obtain genic sequences (Ferguson et al., 2013), and a subset of Pack-MULE-carried genes are expressed and appear to be under selection (Hanada et al., 2009c). However, the mechanisms underlying this preferential acquisition and the functions of the acquired genes remain unclear. Similarly, in mammals, long interspersed nuclear elements and long terminal repeat retroposons are implicated in the generation of recent duplicates (Bailey et al., 2003; She et al., 2008) in a process referred to as segmental duplication (Fig. 2D; Bailey et al., 2002). Note that this is distinct from the original use of segmental duplication in the plant literature, which referred to rearranged genomic regions derived from WGDs (Arabidopsis Genome Initiative, 2000). Although both Pack-MULEs and mammalian segmental duplicates are associated with TEs, it remains unclear if they are generated via similar mechanisms.

Another TE-associated mechanism that generates dispersed duplicates is retroduplication (Fig. 2E; Drouin and Dover, 1990; Brosius, 1991). Here, mRNAs are reverse transcribed into DNA and inserted into the genome. These duplicate genes are referred to as retrogenes. Similar to Pack-MULEs, more retrogenes are found in rice (1,235) than in Arabidopsis (251; Wang et al., 2006; Abdelsamad and Pecinka, 2014), which also may reflect differences in past and current TE activity. In Drosophila melanogaster, retrogenes tend to be derived from genes that are highly expressed in germ-line tissues (Langille and Clark, 2007). However, because the regulatory sequences in the promoter are usually not duplicated, retrogenes were initially considered dead on arrival (Graur et al., 1989). However, there are several examples of functional retrogenes in mammals and D. melanogaster (Kaessmann et al., 2009). There is also evidence that plant retrogenes may be functional. Rice retrogenes have persisted longer than would be expected for nonfunctional elements and are under selection (Wang et al., 2006). In addition, studies in rice and Arabidopsis found that between one-fourth and one-third of retrogenes, respectively, have similar expression patterns to their parental genes (Sakai et al., 2011; Abdelsamad and Pecinka, 2014), and in Arabidopsis specifically, retrogene expression is up-regulated in pollen (Abdelsamad and Pecinka, 2014). However, while the functions of a few retrogenes have been examined in Arabidopsis (Abdelsamad and Pecinka, 2014), the overall functional significance of these duplicates has yet to be fully demonstrated.

Taken together, WGDs and tandem duplications account for the majority of plant duplicates, but TE-based mechanisms and retroduplication also generate a significant number of duplicates. It should be noted that, for example, in Arabidopsis, these mechanisms combined account for approximately 70% of duplicate genes. It remains to be determined if the remaining 30% were generated by some unknown mechanism or if there was a failure in assigning a duplication mechanism either due to the age-related erasure of specific signatures (i.e. synteny, proximity, and repeats) or too-stringent methods used to assign a mechanism.

GENE DUPLICATE LONGEVITY AND PSEUDOGENIZATION

Half-Life of Duplicate Genes

Despite the contribution of multiple duplication mechanisms and the variance in genome size, plant gene content remains relatively similar across land plant species. Considering that at least five rounds of WGD took place in the land plant lineage leading to Arabidopsis (Bowers et al., 2003; Jiao et al., 2011) and assuming that the common ancestor of land plants had approximately 10,000 genes, the number of genes in extant species would be 320,000 even without taking into account other duplication mechanisms. This expected gene number is approximately 10 times higher than the actual gene number in Arabidopsis and indicates extensive gene loss over time. Thus, although some duplicates have survived over millions to hundreds of millions of years, the predominant fate of most duplicates is loss (Li, 1983; Maere et al., 2005; Hanada et al., 2008). The preponderance of duplicates in plant genomes is driven mainly by the high rate of duplications over evolutionary time accompanied by the preferential retention of some duplicates.

How long will a duplicate survive after duplication? Assuming that the mutational process that leads to duplicate loss is stochastic, the longevity of a duplicate gene can be estimated in the form of half-life (i.e. the amount of time for half of the duplicates derived from a single event [e.g. WGD] to be lost; Lynch and Conery,
2000). The genome-wide half-life of Arabidopsis duplicates is estimated to be 17.3 million years (Lynch and Conery, 2003). For example, if a WGD event happened in the Arabidopsis lineage 17.3 MYA, we would expect that approximately 50% of the duplicates from that event will have been lost. As mentioned earlier, there have been multiple WGDs in the Arabidopsis lineage, so we can evaluate the consistency of duplicate half-life by considering these events independently. The most recent α WGD took place approximately 50 to 65 MYA (Bowers et al., 2003; Beilstein et al., 2010), and the observed duplicate survival rate ranges from 13.3% (Blanc et al., 2003) to 16.3% (Maere et al., 2005). Based on this information, the half-life estimate of α WGD duplicates is 17.2 to 24.8 million years. Although this is not far off from the genome-wide estimate of 17.3 million years (Lynch and Conery, 2003), α duplicates in general have a longer half-life than when all duplicates are considered. How about a more ancient WGD? The γ duplication likely took place approximately 140 MYA (Bowers et al., 2003; Moore et al., 2007), and 44% of duplicates are still retained (Maere et al., 2005). This implies a longer half-life, 31.3 million years, compared with α duplicates. On the other hand, certain gene families across angiosperms show bias against having paralogs from older WGD events, suggesting that their half-lives are shorter than average (Li et al., 2016). This difference in estimated half-life between the α and γ WGD events suggests that duplicate longevity is not constant over time; the rate of duplicate loss appears to decrease as the time since duplication increases, perhaps because a greater proportion of older duplicates are retained due to selective constraints.

Based on studies comparing duplicate and singleton (no closely related paralog) genes, the longevity of duplicates may be influenced by molecular and biological functions (Maere et al., 2005; Hanada et al., 2008), structural features (Jiang et al., 2013), number of protein interactions (Makino and McLysaght, 2012), and parent of origin (Song et al., 1995). Additionally, duplication mechanisms have different impacts on gene content. For example, WGD increases gene content dramatically but happens relatively infrequently. In contrast, tandem duplication, although affecting a limited number of genes, can increase or decrease gene number every meiotic division. These types of features that may affect duplicate retention are not taken into account when estimating genome-wide half-life (Lynch and Conery, 2003), and thus significant deviations are expected. Nonetheless, whether certain types of duplicates have longer or shorter half-lives, even the most conservative estimates suggest that most duplicates are lost relatively quickly after they are generated.

Mechanisms of Duplicate Gene Loss

The process of duplicate loss may involve deletion of the entire duplicate sequence and/or pseudogenization through loss-of-function mutations (Fig. 3A). If two duplicate genes are completely identical, there should not be a penalty for deleting either copy. In reality, however, duplicates are rarely equal. For example, copies derived from TE-mediated duplication, retroduplication, or tandem duplication may be missing parts of the parent gene-coding and/or regulatory regions. In these cases, loss of the new duplicate will likely incur no fitness penalty. Even in WGD, particularly when it involves allopolyploidy (merging of two related, but not identical, genomes), the patterns of duplicate loss are far from random (Thomas et al., 2006). The wholesale loss of duplicates is an important feature of fractionation, the reduction of genic content post WGD (Freeling et al., 2012). Analysis of syntenic blocks produced by the α WGD in Arabidopsis revealed that genes in one duplicate block were lost preferentially (Thomas et al., 2006). This fractionation bias has been observed in several plant species, including Gossypium raimondii (diploid cotton; Renny-Byfield et al., 2015), maize (Schnable et al., 2011), and B. rapa (Cheng et al., 2012). Importantly, this bias applies only to gaps in syntenic blocks that span multiple genes but not to gaps resulting from deletions of single genes (Thomas et al., 2006). Simulations of fractionation based on yeast data suggest that the observed bias in gene loss requires deletion events covering multiple genes and that random, single gene losses alone do not explain the observed pattern of fractionation (van Hoek and Hogeweg, 2007). What is the basis for fractionation bias? Looking across multiple WGD events, fractionation appears biased against duplicates with reduced expression level and promoter complexity (Schnable et al., 2012). Supporting this idea, analysis of lowly expressed genes in Arabidopsis suggests that they may be undergoing more rapid divergence and possibly pseudogenization (Yang et al., 2011).

Nonfunctional duplicates are not always deleted; plant genomes are littered with thousands of apparently degenerated, nonfunctional duplicates referred to as pseudogenes (Benovoy and Drouin, 2006; Guo et al., 2009; Zou et al., 2009a). Pseudogenes are identified based on their similarity to annotated genes and the presence of disabling mutations (e.g. premature stop codons and frame shifts in protein-coding genes) that lead to presumed loss of function (Vanin, 1985). Although pseudogenes are presumably nonfunctional, a small subset of pseudogenes in rice and Arabidopsis are clearly expressed (Yamada et al., 2003; Thibaud-Nissen et al., 2009; Zou et al., 2009a). There are three potential explanations for pseudogene expression. First, some pseudogenes may have been falsely predicted due to misannotation. An example is the rice ent-KAUEREN SYNTHASE LIKE2 gene, which had mispredicted coding regions (Tezuka et al., 2015). This issue will likely be less prominent as transcript-based annotations improve (Law et al., 2015). The second explanation is that some pseudogenes may still be functional as truncated proteins or as RNA. For example, apomixis in the grass Paspalum simplex is hypothesized to be the consequence of antisense regulation by the transcript of a
pseudogene related to the ORIGIN RECOGNITION COMPLEX3 gene (Siena et al., 2016). However, there is no direct evidence that the pseudogene is responsible for the antisense transcript. Finally, some pseudogenes may have become pseudogenized relatively recently and are in the process of complete decay. This is
consistent with the finding that expressed pseudogenes tend to be derived from more recent duplication events (Zou et al., 2009a). In addition, pseudogenes, expressed or not, tend to have elevated nonsynonymous (amino acid-changing) substitution rates (Zou et al., 2009a), indicating that they are not subject to the same degree of selective pressure as their intact relatives.

Do duplicates have similar propensities for pseudogenization? In *B. rapa* and *R. raphanistrum*, which experienced a recent genome triplication event, there are significantly more pseudogenes than in the related species Arabidopsis and *Arabidopsis lyrata*, which did not experience a recent WGD (Moghe et al., 2014). Also, there is a significant, positive correlation between gene family size and the number of pseudogenes within families (Zou et al., 2009a). Thus, in general, the more members of a family that are duplicated, the more losses occur. However, this correlation is far from perfect (Zou et al., 2009a), indicating that other factors are important. One such factor is gene function. For example, Arabidopsis pseudogenes tend to have functional relatives playing roles in disease resistance, specialized (secondary) metabolism, cell wall modification, and protein degradation (Zou et al., 2009a), but transcription factor and receptor-like kinase families tend not to have pseudogenes (Hanada et al., 2008). In addition, pseudogenes tend to be derived from tandem duplicates (Hanada et al., 2008), although this may be due to the higher rate of tandem duplication compared with other duplication mechanisms. Taken together, duplicate longevity depends on functional role and duplication mechanism, which necessarily means that there is a significant bias in the kinds of duplicates that are retained.

### MECHANISMS FOR RETENTION OF DUPLICATE GENES

#### Genetic Drift and Genetic Redundancy

Over the course of plant evolution, hundreds of thousands of new genes were created by duplication, and most of these duplicates were lost over time. Nonetheless, considering that more than half of the gene content in most plant species consists of duplicates, some duplicates have clearly escaped this fate. Why do some duplicates persist while others are lost? Models for duplicate gene retention in general have been reviewed elsewhere (Innan and Kondrashov, 2010; Maere and Van de Peer, 2010); here, we will focus on examples of duplicate retention in plants (Fig. 3, B–G). It is important to note that these models are not mutually exclusive. For example, selection on both duplicates to maintain dosage balance (Fig. 3D) contributes to increased duplicate longevity, which may allow time for the evolution of novel functions (Fig. 3E; Veitia et al., 2013). Here, we discuss each model independently to emphasize the distinct mechanisms that contribute to duplicate retention. First, we discuss the idea that both duplicates are retained without a significant change in function, either because insufficient time has passed for deleterious mutations to accumulate or because there is selection pressure to retain redundant functions.

Assuming that mutations accumulate randomly and that selection is not a factor, genetic drift will be the dominant factor influencing the frequency of mutant alleles (Kimura, 1968, 1983). In this case, a mutation that appears in a gene would take approximately four $N_e$ generations to become fixed (where $N_e$ is effective population size; Kimura and Ohta, 1969). In Arabidopsis, which has an estimated $N_e$ of 250,000 (Cao et al., 2011) and a winter annual life cycle for most accessions (Michaels et al., 2004), the time to fixation is approximately 1 million years. When a pair of duplicates is considered, the situation is more complicated because either copy can be lost without affecting fitness. Assuming that the time for a recent WGD to sweep through the population is negligible, the time to fixation is a function of $N_e$, the fitness effect of the loss of both duplicate genes, and the mutation rate at the duplicated loci (Kimura and King, 1979). The average time to fixation is estimated to be between three and 20 $N_e$ generations. In Arabidopsis, this translates to an average fixation time of between 0.75 and 5 million years. Thus, it is expected that some duplicate genes potentially survive for several million years due to genetic drift and not because their presence is beneficial. In this situation, some duplicates may be decaying functionally even though there is no apparent sign of pseudogenization (Lehti-Shiu et al., 2015). However, if drift were the only factor affecting retention and the expected time to deletion was approximately 1 million years, we would expect a much lower genome-wide duplicate half-life (mean lifetime approximately $1.44 \times$ half-life). Thus, a substantial number of duplicates are most likely under selection (i.e. loss of function in either of the duplicates is expected to reduce the fitness of the individual).

Alternatively, duplicate retention might occur via selection for genetic redundancy (or genetic buffering), where the effects of a null mutation are ameliorated (or buffered) due to the presence of an intact, duplicate copy (Zhang, 2012). The prediction is that developmental or physiological phenotypes are only obvious when a gene and one or more of its relatives (paralogs) are mutated (Nowak et al., 1997). Consistent with the idea of genetic redundancy, phenotypic effects when one copy of a duplicated pair is disrupted are significantly smaller compared with those observed when a singleton gene is disrupted in Arabidopsis (Hanada et al., 2009a). However, claims of genetic redundancy between duplicates thus far are based on the absence of gross morphological, developmental, and/or behavioral phenotypes in highly controlled environments. Thus, relatively subtle phenotypic changes or conditional phenotypes resulting from mutations in one duplicate copy, which may have fitness consequences, may not be detected. Additionally, although redundancy may be beneficial in a way that is analogous to a fail-safe in an engineered system...
(McAdams and Arkin, 1999; Kitano, 2004; Kondrashov, 2010), it remains to be shown how selection can act in anticipation of future need to favor redundancy. Although long-term conservation of redundant duplicates is feasible in simulated situations, it requires perfect equivalency in gene functions and in mutation rates between the two duplicates (Nowak et al., 1997), which is highly unlikely. After gene duplication, various degrees of functional redundancy are expected. But the redundant functions can be present not necessarily because they are useful but because there has been insufficient time for their loss. Due to the challenges in experimentally assessing the functional and fitness contributions of duplicates, the true extent of genetic redundancy and whether redundancy is the cause or consequence of duplicate retention remain unclear.

Selection on Existing Functions

Duplicates can be retained without acquiring new functions via one of the following four mechanisms: gene dosage increase (Ohno, 1970), DDC (Force et al., 1999), gene balance (Freeling and Thomas, 2006), and paralog interference (Baker et al., 2013). Ohno (1970) recognized that, in situations where increased gene dosage confers an advantage by meeting metabolic demands, the presence of duplicate copies is beneficial (Fig. 3B). Unlike redundancy, the robustness of duplicates is clearly selectable (McAdams and Arkin, 1999; Kitano, 2004; Kondrashov, 2010); although the molecular function of the duplicates may be unchanged, the effect of increased dosage is new. Models of budding yeast gene networks suggest that WGD likely contributed to increased flux in the glycolytic pathways, which confers a fitness benefit in high-Glc environments (van Hoek and Hogeweg, 2009). Similarly, in Arabidopsis, duplicate retention after WGD is associated with reactions with high metabolic flux (Bekaert et al., 2011), suggesting that increased gene dosage results in increased metabolic activity, which may be beneficial. The impact of duplicate gene dosage is emphasized in a recent review (Conant et al., 2014).

In DDC or subfunctionalization (Fig. 3C; Force et al., 1999), after duplication of a multifunctional gene, each copy randomly loses subfunctions of the original gene (degeneration), and because each duplicate loses different subfunctions, both copies have to be kept to maintain the original, ancestral functionality (complementation). An important point to emphasize is that duplicate retention through DDC does not require any new evolved functions, just partitioning of the old ones. Evidence suggests that DDC has occurred at the level of protein function (Akilíu et al., 2014; Akilíu and Culligan, 2016), at the level of gene expression (Duarte et al., 2006; Throude et al., 2009; Zou et al., 2009a; Ma et al., 2015), and at both levels simultaneously (Geuten et al., 2011). Under DDC, subfunctions are assumed to be independently mutable, and it is expected that the ancestral functions are partitioned symmetrically among duplicates. Interestingly, the pattern of functional partitioning between duplicates tends to be highly asymmetric, where one copy tends to have significantly more subfunctions than its paralog, compared with what would be expected randomly (Zou et al., 2009b). Thus, either subfunctions are not independently mutable and/or other factors affect the partitioning of subfunctions.

An alternative model for duplicate retention is the gene balance hypothesis (Birchler et al., 2005; Birchler and Veitia, 2007, 2010). Under one version of this model, after WGD, duplicate genes that are dosage sensitive tend to be retained (dosage balance; Fig. 3D; Thomas et al., 2006). The idea is that duplication of a gene whose product has a greater number of molecular interactions (e.g. protein-protein or protein-DNA interactions) will lead to a dosage imbalance if all its interactors remain single copy. The individual harboring the duplicate will then have reduced fitness due to this imbalance. A higher degree of imbalance is expected for gene products with a higher number of interactions. In this situation, the duplication of just one gene will likely have a deleterious effect. But when this highly connected gene is duplicated along with its interactors in a WGD event, its retention is favorable because its removal would lead to imbalance. Consistent with gene balance, the expression patterns of highly interconnected genes tend to be more highly correlated than random duplicates (Lemos et al., 2004). Furthermore, there is a greater correlation between the expression patterns of WGD duplicates than between tandem duplicates (Casneuf et al., 2006; Arabidopsis Interactome Mapping Consortium, 2011).

Gene balance also can result from mechanisms other than dosage balance. In situations where molecular interaction is important for gene function, degenerative mutations in one duplicate copy may interfere with the functions of its paralog (paralog interference; Fig. 3E). These degenerative mutations will be selected against, leading to retention of both duplicates (Bridgham et al., 2008; Baker et al., 2013). Paralog interference is distinct from dosage balance in two respects. First, paralog interference occurs at the level of protein interaction and is independent of changes in the amount of gene products. Second, it is relevant specifically to situations where the formation of homomultimers is important for the ancestral gene function. The sequestration of interactors by the degenerate copy effectively creates an imbalance but not in dosage. Given that homomultimerization is prominent in multiple protein families such as transcription factors (Amoutzias et al., 2008), paralog interference may have a significant contribution to duplicate retention post WGD (Kaltenegger and Ober, 2015). Note that paralog interference also is distinct from DDC, which requires degenerative mutations to explain retention.
Contribution of Selection on New Functions to Duplicate Retention

Duplicates, although originally sharing the same functions, may acquire new functions. If these functions are beneficial, selection will act to retain both duplicates. Two models that explain duplicate retention due to the acquisition of novel adaptive functions are neofunctionalization (Ohno, 1970) and EAC (Des Marais and Rausher, 2008). Under the neofunctionalization model, one duplicate retains the ancestral function while its paralog gains a novel function (Fig. 3F). If the novel function contributes to better fitness, selection should maintain both duplicates. Note that determining whether neofunctionalization has taken place requires knowledge of gene functions prior to duplication. Some examples where neofunctionalization after duplication has likely contributed to duplicate retention include MADS box transcription factors involved in the evolution of novel floral structures (He and Saedler, 2005; Hu et al., 2015b; Zhang et al., 2015), 4,5-dioxygenase and cytochrome P450 genes that contribute to pigment variation in Caryophyllales (Brockington et al., 2015), and the recruitment of duplicated primary metabolite genes into specialized metabolite pathways (Durbin et al., 2000). At the gene expression level, it is estimated that approximately 10% of Arabidopsis duplicate genes have gained a novel response to stress conditions (Zou et al., 2009b), although it remains to be determined whether these novel responses are adaptive or not.

Similar to DDC, EAC (Hittinger and Carroll, 2007; Des Marais and Rausher, 2008) predicts subfunctionalization followed by the specialization of duplicates, but in EAC, novel functions arise prior to duplication (Fig. 3G). Subsequent subfunctionalization allows the separate functions of the ancestral gene to evolve independently. Thus, the distinguishing characteristic of EAC is improvement of the original ancestral function in one duplicate and of the novel function in the other (Des Marais and Rausher, 2008). EAC also is similar to the classic neofunctionalization model in that duplication allows further adaptive changes to accumulate, but the neofunctionalization model postulates that, after duplication, only one copy acquires a novel function. Under both neofunctionalization and EAC, duplicates are retained due to their adaptive contribution: individuals with duplicates that have novel and/or optimized functions are expected to have higher fitness compared with individuals with single copies. In contrast, under DDC, the duplicates do not provide an adaptive edge compared with the single-copy gene. Because of the requirement that the novel function existed prior to duplication, EAC is likely particularly relevant to genes that are nonessential and not highly conserved but are highly liable to selection (Sikosek et al., 2012). Examples of EAC in plants include improved enzymatic activity on ancestral flavonoid substrates after duplication of dihydroflavonol-4-reductase leading to adaptive changes in anthocyanin synthesis (Des Marais and Rausher, 2008) and salicylic acid/benzoic acid/theobromine enzymes, where improved enzymatic activities likely are under positive selection (Huang et al., 2012).

The models that we have presented here as distinct may actually involve a combination of different mechanisms. For example, EAC invokes both the neofunctionalization and DDC models to explain duplicate retention. Similarly, under the subneofunctionalization model (He and Zhang, 2005), the expected outcome of duplication is first the partitioning of ancestral functions followed by neofunctionalization in both duplicates. Subneofunctionalization is distinct from EAC, however, in that the total number of unique functions of the duplicate pair is expected to exceed the number of original functions. This is consistent with the finding that the numbers of protein interactions among two duplicate genes are higher than those of randomly paired singletons (He and Zhang, 2005). Paralog interference does not require adaptive change as a mechanism for duplicate retention, but subsequent neofunctionalization may resolve interference in a way that further increases the adaptive value of both duplicates (Baker et al., 2013; Kaltenegger and Ober, 2015). Taken together, the retention of a duplicate may involve any individual or combination of the above mechanisms, and it is of interest to know the relative contributions of each. Although some studies have explicitly compared different models of retention (Yang et al., 2006), the major obstacle to assessing the mechanisms of duplicate retention remains the inference of ancestral function, which requires a number of assumptions and is hypothetical. Thus, assessing the relative contributions of each duplication mechanism, or combination of mechanisms, to duplicate retention remains a major challenge.

PROPERTIES OF RETAINED DUPLICATES

Evolutionary Rate of Duplicates

After gene duplication, the rate of evolution (sequence substitutions) should increase, at least initially, because the presence of two copies relaxes selection against previously deleterious mutations (Scannell and Wolfe, 2008). Consistent with this expectation, duplicates display relaxed purifying selection (Carretero-Paulet and Fares, 2012) as well as differences in sequence structure, such as length of the coding region and the size and distribution of indels (Wang et al., 2013a). This increase in evolutionary rate is not necessarily equal for both copies: the proportion of WGD duplicates with evidence of asymmetric evolution ranges from 21.2% in maize to 68.3% in P. patens (Carretero-Paulet and Fares, 2012). Additionally, after a recent genome triplication approximately 25 MYA, 13% to 19% of B. rapa and R. raphanistrum paralogs evolved asymmetrically (Moghe et al., 2014). This pattern
is expected under the neofunctionalization model, where one copy maintains the original function while mutations that contribute to new functions accumulate in the other copy. However, this pattern also can be explained by the gradual loss of function in one copy: once one degenerative mutation has occurred in one duplicate, the chance for a second mutation to occur in the same duplicate is expected to be higher. Thus, for any particular case of asymmetry, the underlying cause needs to be determined, and neofunctionalization cannot be assumed by default.

Asymmetry in evolutionary rate demonstrates differential evolution between duplicate copies, but what about relative to the previous, unduplicated copy? To assess whether duplication is associated with altered evolutionary rates, three different comparisons can be made. First, a duplicate pair can be compared with the putative ancestral gene. This is exceedingly difficult due to the challenges associated with inferring ancestral function. Second, a duplicate pair can be compared with a closely related singleton (that was duplicated in the past but whose paralogs were lost). In this case, the duplicates and singletons have common ancestry and, presumably, similar functions. An exemplary study of this type provides evidence for asymmetric sequence evolution of both WGD and small-scale duplicates in multiple plant species (Carretero-Paulet and Fares, 2012). Third, all duplicates (genes with paralogs) can be compared with all singletons (those with no obvious paralog) in order to assess the average trend. In Arabidopsis, the nonsynonymous (amino acid-changing) substitution rates of duplicates tend to be lower compared with those of singletons (Yang and Gaut, 2011). The apparently more constrained evolution among duplicate genes can be the consequence of gene balance and/or paralog interference, but there can be other confounding factors that complicate the comparison of duplicates and singletons. For example, the evolutionary rate differences between duplicates and singletons can reflect differences in gene functions (discussed below) and sequence features such as protein domains, which tend to be longer, more numerous, and more highly conserved in duplicates than in singletons (Chapman et al., 2006). Additional sequence features differ between singleton and duplicate genes across multiple plant species (Jiang et al., 2013), but it remains to be seen whether these features contribute to the difference in evolutionary rate or are themselves consequences of duplication.

Expression Patterns of Duplicates

In addition to sequence differences, duplicate genes can have divergent expression patterns. At the transcriptional level, approximately 70% of duplicate pairs in Arabidopsis have significant differences in transcript levels (Ganko et al., 2007). In Gossypium (cotton), the transcript levels of 99.4% of the duplicates derived from a WGD event that occurred 60 MYA have diverged (Renny-Byfield et al., 2014). Expression divergence at the protein level also has been documented (Hu et al., 2015a). This divergence in expression may be under selection due to subfunctionalization and/or neofunctionalization. However, this difference in expression could result from fractionation among WGD duplicates (Schnable et al., 2011) that may not be subject to selection, at least initially. If expression differences were purely neutral, paralogs from younger duplication events would be expected to have more similar expression profiles. Yet the correlation between duplicate expression similarity and the timing of duplication (using synonymous substitution rate as a proxy) is weak, and the direction of correlation is not consistent between WGD (positive) and tandem (negative) duplicates (Haberer et al., 2004; Ganko et al., 2007). This is also true in rice (Li et al., 2009) and poplar (Populus spp.; Rodgers-Melnick et al., 2012). Thus, expression divergence is not an entirely neutral process. In contrast, there is a highly significant negative correlation between nonsynonymous substitution rate and expression similarity between duplicates (Ganko et al., 2007). That is, the more divergent the protein sequences are, the more similar the expression profiles. Assuming that the duplicates that contribute to this pattern are indispensable, it would appear that duplicates will be retained if they have sufficiently large differences in either expression profiles or sequences and, in some cases, a combination of both. Consistent with this notion, younger duplicates that are essential (lethal when mutated) do, in fact, have more divergent expression profiles compared with older, essential duplicate genes (Lloyd et al., 2015). In this case, the large difference in expression patterns of young duplicates contributes to the lack of buffering effect if one duplicate is lost.

Expression divergence between duplicates can arise due to differences at various stages of expression regulation. First, differences in cis-regulatory elements between Arabidopsis duplicate genes explain, in part, why duplicates have different responses to stressful environments (Zou et al., 2009b). In Gossypium hirsutum (allotetraploid cotton), 40% of homologs derived from a recent WGD display transcriptional divergence that can be attributed to cis-regulatory divergence between the diploid progenitors (Chaudhary et al., 2009). This result also is consistent with a recent study that found divergence in DNAseI footprints in the promoters of recently duplicated genes, suggesting that they are regulated by different sets of transcription factors (Arsovski et al., 2015). The degree of regulatory interaction divergence is dependent on the duplication mechanism: most WGD-derived duplicates share some regulatory interactions, while non-WGD duplicates tend not to have overlapping regulators (Arsovski et al., 2015). Gene body methylation is found to impact transcription (Lister et al., 2008; Takuno and Gaut, 2012), and divergence in methylation pattern between duplicates is significantly, although rather weakly, correlated with expression divergence in Arabidopsis (Wang et al., 2014a), rice (Wang et al., 2013b), and cassava...
duplicates also can differ significantly in alternative splicing. For example, in hexaploid bread wheat, 42% and 61% of alternative splicing events differ between homologous gene pairs in chromosome 3A-3B and 3A-3D comparisons, respectively (Akhunov et al., 2013). In Arabidopsis, 85% of a WGD and 89% of tandem duplicates have divergent alternatively spliced forms (Tack et al., 2014).

Much like evolutionary rate, the expression of duplicates as a whole can be significantly different from that of singletons (Holstege et al., 1998; Seoighe and Wolfe, 1999). In plants, duplicates tend to be consistently more highly expressed than singletons in multiple plant species (Jiang et al., 2013). Conversely, while breadth of expression also is higher in duplicates in general, the trend is not universal. The differences in chromatin accessibility between duplicates and singletons may provide an explanation for expression level differences: in Arabidopsis, the promoter regions of duplicate genes have nearly twice as many DNase I hypersensitive sites compared with singleton genes (Arsovski et al., 2015). Taken together, these studies highlight the molecular mechanisms that underlie expression divergence between duplicates and between duplicates and singletons. The challenge now is to distinguish between the expression differences that contribute to differences in duplicate functions and those with no significant impact.

Functions of Duplicates

Given that some, although not all (e.g. gene balance), models of duplicate retention involve selection on gain- or loss-of-function mutations, we expect to find evidence of functional difference in duplicate pairs. For the purposes of this review, we classify functions into two different categories: molecular function and biological process-based function. Molecular function is defined as the molecular activity of a gene product (e.g. protein-protein interaction). Analysis of Arabidopsis interactome data revealed that duplicates tend to have different interaction partners (Carretero-Paulet and Fares, 2012; Guo et al., 2013). Nearly half of all WGD duplicate pairs are completely diverged, with no shared protein-protein interactions (Guo et al., 2013). Younger duplicate pairs tend to have more shared interactions, but a similar portion of young (92.7%) and old (97.3%) duplicates have at least some divergence in protein-protein interactions. Greater discordance in protein-protein interactions tends to be correlated with decreased expression similarity and a greater number of distinct protein domains (Guo et al., 2013). Duplicate genes as a whole encode proteins with significantly more protein-protein interactions compared with singleton genes (Alvarez-Ponce and Fares, 2012). However, while genes derived from WGD have greater regulation connectivity, tandem and WGD duplicates show no significant differences in the number of associated protein-protein interactions (Carretero-Paulet and Fares, 2012).

The function of a gene also can be defined as the impact of its gene product on a biological process. There is a large body of experimental evidence demonstrating how biological process functions differ between two plant duplicates. For example, paralogous MADS domain transcription factors control different aspects of plant development, and this functional divergence is due to changes in both promoter and coding regions (McCarthy et al., 2015). To summarize how the biological processes involving duplicates differ from those involving singleton genes, one approach is to examine annotated gene functions. Importantly, the functional differences between duplicates and singletons are heavily influenced by the duplication mechanism. For example, Arabidopsis WGD duplicates involved in signal transduction and transcriptional regulation tend to be retained (Blanc and Wolfe, 2004b; Seoighe and Gehring, 2004; Maere et al., 2005; Shiu et al., 2005). On the other hand, Arabidopsis and rice genes in Gene Ontology (GO; Ashburner et al., 2000) categories relevant to response to environmental stress tend to be tandemly duplicated (Rizzon et al., 2006; Hanada et al., 2008), and Arabidopsis genes that are transcriptionally responsive to stress have experienced lineage-specific expansion due mainly to tandem duplication (Hanada et al., 2008). To assess if the functional biases in duplicate retention are consistent across species, the enrichment of GO terms in WGD and tandem duplicates was evaluated in six flowering plants (Jiang et al., 2013). The overrepresentation of GO terms was found to be anti-correlated between WGD and tandem duplication within each species. However, only a subset GO terms (transcriptional regulation, ribosomes/translation, amino acid synthesis, and kinase activity) had the same pattern of enrichment in the majority of species (Jiang et al., 2013). This is consistent with another study that found differences between the functions of WGD and tandem genes but little correlation between the overall pattern of GO category enrichment in duplicate genes across four plant species (Carretero-Paulet and Fares, 2012). Conversely, gene families that are common across sequenced angiosperms are biased toward single-copy status and retained as such in all species except those with a recent WGD (less than 50 MYA; Li et al., 2016). Although GO-based analysis has been informative, there can be significant issues with its use (Rhee et al., 2008), particularly because there are differences in the quality of functional annotation for different genomes. Further meta-analysis using compiled experimental data across species, which is routinely done in evo-devo studies of floral evolution, for example (Fornoni et al., 2016), may provide the required resolution and accuracy to assess functional differences between duplicates and singletons.
Taken together, duplicate and singleton genes have significantly different sequence properties, expression patterns, molecular functions, and biological roles. These properties can be used to construct quantitative models that can predict whether a duplicate will be retained or not. Among the properties that distinguish duplicates from singletons, which ones are more important? Based on duplicate genes from six flowering plant species, the number of unique protein domains (another metric for the number of functions), nucleotide diversity, and GC content were found to be the most important predictors (Jiang et al., 2013). In another study, where multiple sequence properties, measures of conservation, molecular functions, and biological network information were consolidated in a single quantitative model to predict retention (Moghe et al., 2014), it was shown that no single property could accurately distinguish between duplicates and singletons. However, models that used only GO terms, sequence, and conservation features could predict duplicates equally well as models including the full set of features (Moghe et al., 2014), highlighting the importance of these features in duplicate retention.

EVOLUTIONARY, ECOLOGICAL, AND AGRONOMIC IMPACTS OF GENE DUPLICATION

Contribution of Duplicate Genes to Evolutionary Novelty

A common theme among models of duplicate retention is that, for both copies of a gene to be maintained, differences in function, expression, or interaction occur in one if not both paralogs. In some cases, duplicates will acquire novel functions and contribute to evolutionary novelties. For certain duplicates, evidence for a novel function can be seen in the morphological phenotypes caused when they are knocked out, and the presence of such a phenotype is correlated with the extent of sequence and expression divergence (Hanada et al., 2009b). The impact of duplication on evolution has been discussed in depth (Van de Peer et al., 2009b). Here, we focus on plant examples and classify evolutionary novelties into three types: (1) novel molecular function (e.g. expression in a novel context or interaction with a new protein); (2) novel plant structure/function that results from a new molecular function (e.g. a new floral organ or a new cell type); and (3) novel adaptive traits that result from a novel structure/function (e.g. disease resistance). This distinction is important because the novel functions/structures in types 1 and 2 need not be adaptive and, thus, may be novel without being selected for. What is the evidence that gene duplication has contributed to these three types of novelties in plants?

Novel molecular functions acquired after gene duplication can easily be seen in the context of gene expression. For example, Arabidopsis duplicates have likely accumulated novel environmental responses (Zou et al., 2009b) and novel developmental regulatory patterns (Liu et al., 2011). Novel leaf gene expression patterns also are observed for some maize duplicate genes (Hughes et al., 2014). However, a novel expression pattern is not necessarily adaptive. In the case of novel environmental responses, some of these responses may be adaptive (Zou et al., 2009b), but direct evidence is not available. Duplication also has contributed to the generation of novel metabolic functions. Specialized metabolism genes are retained preferentially following tandem duplication (Yu et al., 2015), and tandemly duplicated clusters have been found in multiple plant species (Kliebenstein and Osbourn, 2012; Nützmann and Osbourn, 2014). Interestingly, some specialized metabolic genes likely arose from neo-functionalized duplicates of primary metabolism genes (Qi et al., 2006), and further duplications of specialized metabolic genes have likely contributed to additional novel biochemical activities (Field and Osbourn, 2008; Takos et al., 2011). In some cases, these duplicates are implicated in defense against herbivores and microbes as well as in attracting pollinators (Mizutani and Ohta, 2010; Moghe and Last, 2015). However, the adaptive significance of most novel biochemical activities has yet to be demonstrated.

There also is evidence that duplicate genes have contributed to novel plant structures/functions, and in some cases, there is clear adaptive significance. A key example is the specification of floral organ identity by MADS box transcription factors (Bowman et al., 1989; Sommer et al., 1990; Coen and Meyerowitz, 1991; Theissen, 2001). Although the MADS box gene family originated prior to the divergence between major eukaryotic lineages (Graham et al., 2010), extensive duplication events have resulted in the expansion of MADS box genes involved in floral development (Purugganan et al., 1995; Kramer et al., 1998, 2006; Causier et al., 2010). Many MADS box paralogs appear to have redundant functions (Vandenbussche et al., 2004; de Martino et al., 2006; Geuten and Irish, 2010), but others have clearly diverged in function (Airoldi and Davies, 2012). Some examples include the divergence of the AP3/TM6 paralogs in petunia (Petunia hybrida) due to a change in regulatory elements (Rijkkema et al., 2006), the modification of preexisting floral regulators leading to the development of novel floral organ types in the Ranunculales (Kramer et al., 2007; Rasmussen et al., 2009), and the lineage-specific expansion and divergence of the DEFICIENS genes in orchids (Orchidaceae spp.), giving rise to unique floral features (Mondragón-Palomino and Theissen, 2011). These examples highlight the importance of duplication and divergence of floral identity genes as a major contributor to the evolution of floral morphology. This process also has likely been important to the evolution of vegetative organs. For example, a duplicated KNOX transcription factor has acquired a novel regulatory pattern that regulates leaf shape and aboveground architecture in plants (Furumizu et al., 2015). Finally, duplication can contribute to the interactions of plants with other organisms: the duplication of a receptor-like...
kinase gene originally involved in mycorrhizal symbiosis gave rise to the lysin motif receptor-like kinase SILYK10 in tomato (*Solanium lycopersicum*), which likely adopted a new role in nodulation with clear adaptive significance (Buendia et al., 2016). These are but a few examples of evolutionary novelties that can be found in the large body of plant genomic and functional studies.

It should be emphasized that a novel molecular function may not necessarily contribute to a novel plant structure/function. In addition, a novel plant structure/function may not necessarily be adaptive. The challenge lies not only in determining the presence of a novel activity but also assessing its adaptive significance. Using the disease resistance (*R*) genes as an example, tandem duplication (Rizzon et al., 2006; Yu et al., 2015) and WGD (Cannon et al., 2002; Plocek et al., 2004; Zhang et al., 2014) have contributed to a net gain of *R* genes over the course of plant evolution. Based on detailed functional studies of selected *R* genes, they are important for eliciting proper host defense responses; thus, their presence is clearly adaptive (Dangl and Jones, 2001; Jones and Dangl, 2006). In addition, strong positive selection driving sequence divergence between some *R* gene duplicates has been noted (Bakker et al., 2006; Ratnaparkhe et al., 2011). Thus, the proliferation of *R* genes has been hypothesized to be an indication of their adaptive value (Holub, 2001).

A substantial number of *R* genes may not be under selection, and for those that are, assessing the adaptive significance is challenging because the biotic factors that interact with the *R* gene products are largely unknown. This challenge is not specific to *R* genes but applies to any duplicate gene thought to contribute to increased fitness.

**Ecological Impacts of Duplicate Genes**

The evolutionary innovations ascribed to duplicate genes can have important ecological implications. Gene duplication has contributed to developmental novelties that facilitate interactions between plants and other species. For example, the evolution of floral characteristics is strongly associated with functional groups of pollinators (Fenster et al., 2004). The concept of a pollinator syndrome suggests that selective pressure exerted by pollinators results in the convergent evolution of a common set of floral traits. In orchids, the majority of species have only a single pollinator (Tremblay, 1992), but the evolution of specialized morphology for certain pollinators has occurred multiple times (Johnson et al., 1998). According to the orchid code hypothesis, the proximal cause of the diversity of orchid floral structures can be attributed to two DEFICIENS-like transcription factor duplication events followed by gain and/or loss of function in different orchid lineages (Mondragón-Palomino and Theissen, 2008, 2009). Similarly, the duplication and subsequent diversification of CYCLOIDEA2-like transcription factors in Malpighiaceae is thought to have been important for the evolution of bilateral symmetry (zygomorphic) in flowers from radial symmetry (actinomorphic; Zhang et al., 2010).

In addition to developmental novelties that facilitate ecological interactions, duplicate genes are important components in the evolutionary arms race between plants and pathogens/herbivores. Plant defense against herbivores involves specialized metabolites such as the glucosinolates in Brassicales (Demain and Fang, 2000; Halkier and Gershenzon, 2006) and novel glucosinolate pathway components, which likely derived from duplication events (Edger et al., 2015). In particular, the retention of the core Trp pathway gene duplicates derived from the β WGD event appears to have led to the evolution of the glucosinolate pathway (Edger et al., 2015). Similarly, the previously noted function and expansion of *R* gene families are suggestive of the importance of *R* gene duplication in the interactions between plants and pathogens.

Interestingly, aside from plant-nonplant interactions, both *R* genes and specialized metabolic genes also can impact plant speciation. Speciation results from barriers in gene flow, which can arise at different times during development and can involve many different mechanisms, from molecular to environmental (Rieseberg and Willis, 2007). Based on the genetic level, incompatibility between related species can be explained by the Bateson-Dobzhansky-Muller model, which involves negative epistasis between two or more divergent loci in hybrids (Örr, 1996). After gene duplication, differential loss of function or patterns of subfunctionalization between duplicate loci can, in some situations, result in net loss of function for certain gamete combinations (Force and Lynch, 2000).

Multiple examples of negative interactions resulting from copy number variation and reciprocal silencing of duplicate genes were observed in a survey of speciation genes (Rieseberg and Blackman, 2010). For example, *R* gene duplicates have been hypothesized to play an important role in speciation due to their ability to induce necrosis in hybrids (Bombles and Weigel, 2007). Defensive compounds such as glucosinolates may function as part of a pollinator syndrome (Demain and Fang, 2000; Halkier and Gershenzon, 2006), which can serve as a reproductive barrier. Hybrid incompatibility has been shown to result from the differential expression of tandemly duplicated receptor-like kinases that are involved in innate immunity in Arabidopsis (Smith et al., 2011) and by negative interactions between tandemly duplicated clusters of receptor-like kinases and subtilisin-like proteases in wild rice (*Oryza rufipogon*) and domesticated rice (Chen et al., 2014).

Because a change in ploidy levels is expected to result in instant reproductive isolation, WGD is seen as a major mechanism of plant speciation (Ramsey and
Schemske, 1998). Consistent with this expectation, there is a significant correlation between the presence of a recent WGD event and the number of extant species in Brassicaceae, Cleomaceae, Fabaceae, Poaceae, and Solanaceae (Soltis et al., 2009). Similarly, an estimated 15% of speciation events in angiosperms and 31% in ferns were associated with an increase in ploidy (Wood et al., 2009). Duplication also has been associated with the evolution of interspecies interaction in plants (Edger et al., 2015) as well as the evolution of novel structures/functions in angiosperms (Soltis and Soltis, 2016), which are thought to be important to speciation and diversification. The radiation of plant species appears to lag significantly behind WGDs, which has led to the proposition that WGD and the subsequent development of novel traits primes a population for speciation by a subsequent dispersal event (Schranz et al., 2012; Tank et al., 2015). Consistent with this model, the timing of recent ploidy events in angiosperms is clustered around the Cretaceous-Paleogene extinction (Vanneste et al., 2014), and WGD events and speciation events in conifers occurred around the time of the Permian-Triassic extinction (Lu et al., 2014; Li et al., 2015b). However, while the lag-modulated association between WGD and diversification has been shown to be significant (Tank et al., 2015), the connections between WGD and species radiation are correlational, not cause and effect. Neopolyploid lineages are more prone to extinction than diploids (Mayrose et al., 2011), which is consistent with the observation that the rate of neo-polyploidization in plants is high yet detectable speciation events due to WGD are relatively rare (Otto and Whitton, 2000; Ramsey and Schemske, 2002; Jiao et al., 2011; Moghe and Shiu, 2014). Given that speciation by WGD is rare, the correlation between WGD and radiating events may be because they are unlikely to be observed in association with dispersed events.

Taken together, gene duplications, both small and large scale, may be important in plant adaptation to variable abiotic and biotic environments, in speciation, and in contributing to the diversity of angiosperm species. Although there is indirect evidence, direct demonstration of the role of duplicate genes in ecological adaptation is challenging to come by. In a landmark study examining the genetic basis of how plants adapt to their local environment, 15 fitness quantitative trait loci were identified using an Arabidopsis recombinant inbred population derived from a Swedish accession and an Italian accession (Ågren et al., 2013). Interestingly, one major quantitative trait locus contains the C-REPEAT-BINDING FACTOR (CBF) locus, which has three CBF genes known to control plant cold tolerance (Gehan et al., 2013). In the Italian population, CBF2 is nonfunctional, resulting in reduced cold tolerance. Although this is not an example of a new duplicate attaining a novel, adaptive function, it provides direct evidence of the importance of duplicate genes in influencing species range due to abiotic constraints.

Contribution of Duplicate Genes to Agronomic Traits

Gene duplication facilitates the evolution of novel traits that can be subjected not only to natural selection but also to artificial selection, which is important to crop improvement. A recent review has summarized studies revealing that duplicate genes derived from polyploidy can be key to crop domestication and the evolution of stress resistance/tolerance traits (Renny-Byfield and Wendel, 2014). Many crop plants have experienced relatively recent WGDs: 1.5 MYA in G. hirsutum (Li et al., 2015a), 13 MYA in soybean (Schmutz et al., 2010), and 0.5 MYA and approximately 3.5 MYA in wheat (Brenchley et al., 2012). In bread wheat, polyploidy contributed to the grain free-threshing character controlled by complex interactions between the duplicate Q transcription factors (Zhang et al., 2011) and to the soft grain character controlled by the Hardness locus (Chantret et al., 2005). It is also hypothesized that polyploidy has contributed to the expansion of wheat storage protein genes (Brenchley et al., 2012). In G. hirsutum, the recent WGD has resulted in up-regulation and increased selection of fiber genes on the A(t) sub-genome compared with the progenitor genes, which is correlated with the production of longer, spinnable fibers (Li et al., 2015a). Similarly, in soybean, nodulation and oil production gene duplicates tend to be retained post WGD (Schmutz et al., 2010) and may contribute to soybean domestication traits. Based on comparative genomic evidence, it is also argued that selection in Brassica napus has led to the preservation of duplicate oil biosynthesis genes (Chalhoub et al., 2014). Although not all of the above examples provide direct evidence, they are suggestive of the impact of gene duplication, particularly polyploidy, on agronomically relevant traits.

In addition to WGD, smaller scale duplications can contribute to agronomic traits. Given that a number of genes involved in plant stress tolerance or resistance are found in tandem clusters and display copy number variation (Ellis et al., 2000), it is expected that tandem duplication has played a major role in these traits. For example, the NBS-LRR gene family, whose members have roles in biotic stress resistance, are highly variable between rice cultivars (Yang et al., 2008). In terms of abiotic stress tolerance, one prominent example is the rice Submergence1 (Sub1) tandem cluster, which contains multiple ethylene-responsive factor genes (Sub1A, Sub1B, and Sub1C) and is involved in submergence tolerance (Fukao et al., 2006). Through comparisons of rice cultivars and wild rice species, it was shown that the Sub1A haplotype likely arose recently, potentially after rice domestication (Fukao et al., 2009).

Tandem duplicates also are implicated in the evolution of other agronomic traits. For example, copy number variation in a tandem duplicated region controls rice grain size diversity, an important quality trait (Wang et al., 2015). This region contains two copies of Grain Length on Chromosome 7 (GL7) that are homologs of the Arabidopsis LONGIFOLIA gene, which regulates...
cell elongation. Tandem duplication leads to elevated levels of GL7 expression and increased grain length (Wang et al., 2015b). Based on comparisons of wild and domesticated pepper (Capsicum annuum) species, tandem duplicate genes involved in capsaicin biosynthesis have likely contributed to the diversification of pungency in peppers (Qin et al., 2014), which may be the basis of pungency variation among domesticated pepper varieties. Additionally, domesticated tomato provides an example where an agronomic trait was influenced by transposon-mediated duplication. Fruit shape difference between domesticated and wild (Solanum pimpinellifolium) tomato is due to the increased expression of IQD12 after retrotransposon-mediated duplication to a new genomic region (Xiao et al., 2008).

In the previous sections, we have discussed several domestication-related traits that can be attributed to gene duplications. One obvious omission is flowering time. When plants are grown in new environments, the change in photoperiod makes selection for proper flowering time essential. In sunflower (Helianthus annuus), the expression of Flowering Locus T (FT) duplicates is central to the control of flowering time (Blackman et al., 2010). During sunflower domestication, a dominant negative mutation likely occurred in the FT1 duplicate and contributed to delayed flowering. Another example is the Heading Date1 (HD1) locus, which encodes a member of the CONSTANS transcription factor family (Liu et al., 2015). In sorghum (Sorghum bicolor), foxtail millet (Setaria italica), and rice, mutant HD1 orthologs have contributed to delayed flowering time; however, the presence of distinct mutations in each lineage suggests independent origins over the course of parallel domestication (Liu et al., 2015). In summary, because of their abundance and the potential for functional divergence and the acquisition of new functions, duplicate genes have contributed to the evolution of morphological, nutritional, and physiological traits in crops.

**FUTURE DIRECTIONS**

Studies based on accumulating comparative and functional genomic data have contributed to our understanding of the life cycle of duplicated genes, including their origins, longevity, mechanisms of retention, molecular functional implications, and impacts on plant evolution and ecology. Nonetheless, there are still many unresolved questions (see “Outstanding Questions”). Although metrics like half-life provide a good indicator of the average behavior of duplicate genes, there is large variance in duplicate longevity (Lynch and Conery, 2003; Maere et al., 2005). Why are some duplicates retained longer than predicted by the average half-life? What are the factors contributing to short-lived duplicates? The answers to these questions lie in a better understanding of retention mechanisms, including neofunctionalization, dosage effect, EAC, DDC, gene balance, and paralog interference. In particular, the challenge is to determine the relative contributions of these retention mechanisms. Another challenge is that knowledge of functional divergence alone is insufficient to distinguish between retention mechanisms. Knowledge of ancestral functions and expression states, which can only be inferred, also is required.

The retention mechanisms outlined above all involve natural selection on existing and/or novel functions. Thus, there must be a fitness cost if one of the duplicates is lost. There are but a few studies that have directly addressed the fitness contribution of duplicates (DeLuna et al., 2008; Qian and Zhang, 2014). Instead, the great majority of studies on plant duplicate genes have focused on morphological, developmental, and/or physiological phenotypes of loss-of-function mutants in highly controlled environments. In many cases, the lack of a phenotype when one duplicate is lost is attributed to genetic redundancy (Hanada et al., 2009a). Although genetic redundancy is an authentic phenomenon, one cannot rule out the possibility that the specific environment requiring the function of the apparently redundant duplicate has not yet been identified. It is also possible that there are subtle phenotypes that remain undetected but have significant fitness consequences (Ågrena et al., 2013). Testing these two possibilities requires assessing the fitness cost of loss-of-function mutations in the field, preferably in native environments. Recent studies of Arabidopsis local adaptation show that it is feasible to detect minute fitness effects (Ågrena et al., 2013). In addition, especially for recently duplicated genes (i.e., from duplication events that occurred approximately 1 MYA), one cannot rule out the possibility that some duplicates persist because not enough time has passed for them to be removed by genetic drift, even though they are no longer functional (Kimura and Ohta, 1969). This is particularly true for selfing plants like Arabidopsis, where even deleterious alleles are not efficiently eliminated (Bustamante et al., 2002). Given the diversity in life histories and environments encountered by different plant species, knowledge of the past history (i.e., history of selection, bottlenecks, gene flow, changes in effective population size, and mating system) will be necessary to better estimate the contribution of drift to the persistence of duplicates. This knowledge can be partially acquired from comparative studies of variation in duplicate gene content within and between related plant species. Initiatives like the Arabidopsis 1001 Genomes Project (Cao et al., 2011) and parallel efforts in other species may soon provide some insights in this regard.

 Genome-wide studies of duplicate genes have revealed that retained duplicates tend to have particular patterns at the sequence, expression, and molecular function levels (Ganko et al., 2007; Carretero-Paulet and Fares, 2012; Jiang et al., 2013). However, each pattern can only marginally predict duplicate retention (Jiang et al., 2013; Moghe et al., 2014), and there remains much unexplained variance even when most factors that have been shown to be correlated with retention are combined...
in a single predictive framework (jiang et al., 2013; moghe et al., 2014). This suggests that additional, unknown factors may contribute to duplicate retention. Considering that the function of a gene is directly or indirectly influenced by many other genes, the retention of a duplicate gene should be influenced by other genes that are closely linked to the duplicate in the gene network (i.e. in the same functional module), a possibility consistent with the expected outcome of the gene-balance model (birchler et al., 2005; birchler and veitia, 2007, 2010). This network idea can be pushed to an even higher level of organization by asking how other modules influence the retention of duplicate genes in an entire module. Thus, a systems-level understanding (i.e. knowledge of the architecture of the gene networks and the nature of the connections between genes) is essential to assess how duplicates are influenced by other network components. This knowledge will be helpful not only for addressing the rather esoteric question of how duplicate genes are retained but also for understanding how duplicate genes collectively influence molecular functions, physiology, and development in plants. For example, due to the high rate of transcription factor retention, a gene is not only regulated by transcription factors from different families but also bound by multiple members of the same family (macneil and walhout, 2011). The expression level of the gene in question is thus determined by a large number of duplicate factors. Without a systems-level understanding of how duplicates differ in their functions, it will not be possible to gain a complete picture of how a gene is regulated.

Ultimately, our interest in studying duplicate genes lies in their evolutionary, ecological, and agronomic impacts. Although the acquisition of novel molecular functions among duplicates is common (blanc and wolfe, 2004b), this alone is not sufficient to conclude that there was an impact on evolution. An apparently novel function (e.g. a new biochemical activity, pattern of expression, or interaction) with no effect on fitness could have been fixed by genetic drift. When a duplicate pair is retained because of dosage sensitivity and paralog interference, an unrelated change in function may be misidentified as a functional novelty. Furthermore, apparent novelty at the single gene level may result from larger scale changes in the genome following WGD, such as fractionation (schnable et al., 2011). To be able to claim that natural selection is important, either a direct study of fitness or molecular evidence of a nonneutral mutation is required. Similarly, the contribution of duplication to the diversification of a multitude of plant traits has been explored, but few have examined the adaptive significance of those traits (Agren et al., 2013; gehan et al., 2015). More examples illustrating the impact of novel traits on plant adaptation are needed.

Understanding the impact of duplicated genes is important in light of the challenges facing agriculture in the 21st century, including both the old problems of yield, disease resistance, and stress tolerance as well as new issues related to global climate change. Addressing the grand challenge of food security will not only require improving our ability to modify plant traits (halpin, 2005) but also our ability to identify the causal loci of desirable traits (mickelbart et al., 2015) and the genomic context in which they exist (vaughan et al., 2007). In this regard, a continuing effort to understand how duplicate genes have contributed to novel functions, expansion of gene families, and the structure of the genome as a whole is necessary. Considering how duplicate genes have contributed to evolutionary novelties and diversity in plants, understanding the evolution of duplicate gene functions holds the key to understanding the future of both natural and domesticated populations, particularly in light of impending environmental shift due to global climate change.

LITERATURE CITED


Ferguson AA, Zhao D, Jiang N

Dehal P, Boore JL

Panchy et al.

D (2008) Escape from adaptive con

Des Marais DL, Rausher MD


He C, Saedler H (2005) Heterotrophic expression of MPF2 is the key to the evolutionary Chinese lantern of Physalis, a morphological novelty in Solanaceae. Proc Natl Acad Sci USA 102: 5779–5784


Li Z, Banjara AE, Sessa EB, Scascitelli M, Graham SW, Rieseberg LH, Barker MS (2015b) Early genome duplications in conifers and other seed plants. Sci Adv 1: e1501084
McCarthy EW, Mohamed A, Litt A (2015) Functional divergence of APETALA1 and FRUITFULL is due to changes in both regulation and coding sequence. Front Plant Sci 6: 1076


