Duplications and Turnover in Plant Genomes

11

Michael S. Barker, Gregory J. Baute, and Shao-Lun Liu

Contents

11.1 **Introduction** 155

11.2 **Frequency and Sources of Duplications in Plant Genomes** 155

11.2.1 Small-Scale Duplications: Segmental Duplication 157

11.2.2 Large-Scale Duplications: Polyploidy 158

11.3 **Differential Fates of Duplicated Genes in Plant Genomes** 158

11.3.1 Introduction and Nonfunctionalization 158

11.3.2 Neofunctionalization 159

11.3.3 Subfunctionalization 160

11.3.4 Frequency of Subfunctionalization and Neofunctionalization on a Genome-Wide Scale 161

11.3.5 Gene Dosage Balance 163

11.4 **Duplication, Turnover, and Re-organization of Plant Genomes** 165

11.4.1 Re-organization of Polyploid Plant Genomes 166

References 167

11.1 Introduction

Cytologists have long documented differences in chromosome number and organization among plants, but the truly dynamic nature of plant genome evolution is only becoming apparent with fully sequenced and assembled genomes. A major result of these new data is that duplication—of single genes, chromosomes, and whole genomes—is a major force in the evolution of plant genome structure and content. For example, genomic comparisons among divergent animals are able to recover significant signatures of synteny (Hiller et al. 2004), but less divergent flowering plant genomes often demonstrate relatively lower large-scale collinearity because of cycles of polyploidy and diploidization (Tang et al. 2008; Salse et al. 2009). Among individuals and closely related species, copy number variation and changes in gene family size are now recognized as critical sources of genetic variation (Lynch 2007). Duplication and subsequent resolution have yielded a continually changing genome whose elements are constantly turning-over. Although gene and genome duplication has long garnered attention as a potentially important source of evolutionary novelty (Haldane 1933; Stebbins 1950; Ohno 1970), the perspective of a dynamic plant genome fueled by duplication and loss stands in contrast to classical concepts of a largely stable genome. In this chapter we provide an overview of how duplication-driven genomic turn-over has influenced the evolution and diversity of plant genomes.

11.2 Frequency and Sources of Duplications in Plant Genomes

Plant and other eukaryotic genomes contain an abundance of duplications with a variety of origins. The power of these duplications to influence evolution is attenuated by the rate and nature of gene birth and death. In most eukaryotes, including plants, plots of the age of all gene duplications in

M.S. Barker (✉)
Department of Ecology & Evolutionary Biology, University of Arizona, Tucson, AZ, USA
e-mail: msbarker@email.arizona.edu

J.F. Wendel et al. (eds.), *Plant Genome Diversity Volume 1*,
DOI 10.1007/978-3-7091-1130-7_11,

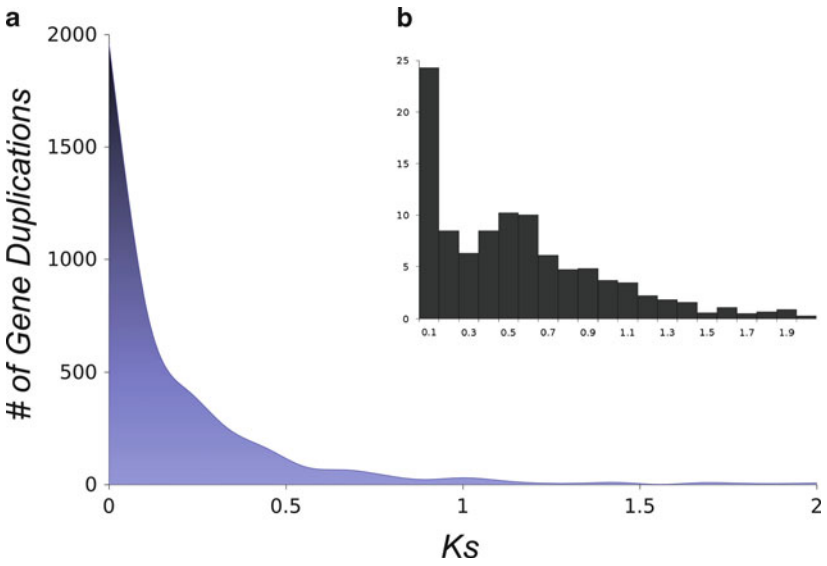

Fig. 11.1 Age distribution of gene duplications from multiple nuclear gene family phylogenies. (**a**) *Selaginella moellendorffii* (Banks et al. 2011), the only sequenced plant genome without an ancient polyploidy. (**b**) *Helianthus argophyllus* (Barker et al. 2008) with a paleopolyploidy visible as a large peak of gene duplications

a single genome demonstrate an exponential or power-law distribution, generally with a steep curve indicative of high birth and death rates (Fig. 11.1). An immediate result apparent from these types of plots is that most duplicated genes do not survive very long. However, polyploidy, or whole genome duplication, yields a large and enduring signature in such plots (Fig. 11.1b). Although these polyploid peaks are observable even after many millions of years because of the scale of gene birth, the signal likely persists longer than expected because some gene classes are known to be strongly retained in duplicate following polyploidy (Maere et al. 2005; Doyle et al. 2008; Edger and Pires 2009). Thus, the nature of gene birth has a significant impact on the ultimate fate of duplicates.

How often does a gene duplicate in the average plant genome? Lynch and Conery (2003) measured the birth rate (B) of genes in *Arabidopsis thaliana* to be 0.0032 with a death rate (D) of 0.033 on a time scale of 1% synonymous divergence per gene for a single individual. Using an alternative approach, Maere et al. (2005) arrived at a nearly identical B for *Arabidopsis*—0.03 per 10% synonymous divergence. Although these estimates are based on the initial peak of duplications and should not be confounded by paleopolyploidy, it is possible that the presence of paleopolyploid duplicates influences the birth and death rates of these young, ongoing duplications in plant genomes. To evaluate this possibility, we used Lynch and Connery's (2003) method to estimate birth and death rates for the recently sequenced genome of *Selaginella moellendorffii* (Banks et al. 2011), the only sequenced plant without evidence of ancient polyploidy. We estimate $B = 0.00883$ and $D = 0.0999$ for *S. moellendorffii* on a scale of 1% synonymous divergence. Although these rates are higher than *Arabidopsis*, the B/D ratios (*Arabidopsis* $B/D = 0.0970$; *Selaginella* $B/D = 0.0883$) are similar, suggesting that paleopolyploidy likely has a negligible impact on the birth and death rates of ongoing gene duplication in plant genomes. Notably, these B/D ratios are the highest estimated rates among sampled eukaryotes, but still of the same order of magnitude (Lynch and Conery 2003). Overall birth rates in most plant genomes will be much higher than these estimates from the initial slope of the birth-death curve because of ancient polyploidy.

To put these rates of gene birth and death into perspective, it is useful to compare them to other evolutionary events. For example, the average gene birth rate for eukaryotes is approximately 40% of the rate of mutation per nucleotide site (Lynch 2007). *Arabidopsis* and *Selaginella* are within the reported range with rates of gene duplication—excluding polyploidy—of nearly 32% and 88% the rate of mutation per nucleotide site. If a whole genome duplication occurs at least once per 100% synonymous divergence—a reasonable estimate for most angiosperms (Blanc and Wolfe 2004a, b; Pfeil et al. 2005; Cui et al. 2006; Barker et al. 2008; Tang et al. 2008; Shi et al. 2010)—then the rate of duplication for plant genes will often exceed the mutation rate over the time scale it takes to achieve mutational saturation. Considering that some whole genome duplications are actually triplications (Jaillon et al. 2007) or possibly higher multiples, and that many plant genomes have experienced more than one round of polyploidy on a time scale of 100% synonymous divergence (e.g., Cui et al. 2006; Barker et al. 2008; Shi et al. 2010), the genome-wide rate of gene duplication is substantially higher than the rate of mutation per nucleotide in plant genomes on a long time scale.

11.2.1 Small-Scale Duplications: Segmental Duplication

Both large and small-scale duplication events produce new genes in plant genomes. Although large-scale duplication events such as whole genome duplications have received significant attention among plant biologists, much more frequent but small-scale duplication events have contributed similar quantities of paralogs. Nearly 25% of the genes in *Arabidopsis* are the product of ancient whole genome duplications (Blanc et al. 2003), but nearly 16% of the genes occur as tandem duplicates (Rizzon et al. 2006), just one class of small-scale duplications. Lynch (2007) describes four mechanisms for the origin of gene-sized duplications in eukaryotic genomes that are collectively referred to as segmental duplications. These include tandemly duplicated genes derived from replication errors such as strand slippage or unequal crossing-over, the replication and ectopic insertion of genes downstream from sloppily transcribed non-LTR retrotransposons, the capture and incorporation of new genes from DNA inserts during repair of double-strand breaks, and duplication by ectopic exchange when the ends of a double-strand break invade a non-homologous region and initiates a recombination event that copies a gene(s) by strand extension before repair of the broken chromosome.

Analyses of plant genomes indicate that these mechanisms actively and variously contribute to the repertoire of gene families across the plant phylogeny. For example, unequal crossing over appears to have been responsible for a large fraction of tandemly arrayed genes in *Arabidopsis thaliana* and *Oryza sativa* (Rizzon et al. 2006). Consistent with duplication by unequal crossing over, Rizzon and colleagues (2006) found nearly 80% and 88% of the tandemly arrayed genes in rice and *Arabidopsis*, respectively, are in direct orientation. Other mechanisms, such as intrachromosomal recombination between repeats, generally produce tandem duplicates that are oriented in opposite directions (Schuermann et al. 2005). Thus, unequal crossing over appears to be responsible for ~14% and ~11% of the genes in the *Arabidopsis* and rice genomes, respectively.

Other mechanisms of segmental duplication have also produced abundant duplications in plant genomes. In particular, repetitive DNA constitutes a large fraction of plant genomes and is an important mechanism of genome size evolution (Leitch et al. 2005). Much of this repetitive DNA is non-coding and derived from a variety of mobile genetic elements. However, many genes or at least gene fragments may be duplicated along with mobile elements. For example, *mutator-like* elements (MULEs) are associated with the transduplication of more than 1,300 genes or gene fragments in rice (Jiang et al. 2004; Juretic et al. 2005). Although most transduplicated gene fragments are pseudogenes (Juretic et al. 2005), some transduplicated genes may be active and contribute to phenotypic evolution (Jiang et al. 2004). An outstanding case is the retrotransposon mediated gene duplication of a *SUN* gene in a cultivar of *Solanum lycopersicum*. *SUN* is a major gene involved in the control of fruit shape, particularly length, in tomato. By mapping and positionally cloning the *SUN* locus, Xiao and colleagues (2008) discovered that the region, which encoded a member of a IQ67 domain-containing family, arose by duplication of a 24.7 kb fragment mediated by the *Rider* long terminal repeat retrotransposon. Critically, duplication and movement of this gene led to increased *SUN* expression that significantly altered the phenotype. Considering the frequency of repetitive elements in plant genomes, their potential contributions to both gene and gene expression evolution may be significant.

Functionally transduplicated gene families may also expand selfishly and contribute little to the host organism. Hoen and colleagues (2006) recently found evidence for functional transduplication and expansion of a family of *ULP*-like genes in *Arabidopsis* associated with MULEs. Unlike most functional cellular gene families, these *ULP*-like genes were targeted by small RNAs and silenced in most tissues but with elevated expression in the pollen. Further analyses indicate that these genes may actually encode for proteins that disrupt host transposon-silencing mechanisms. Thus, this gene family may be selfish rather than contributing to cellular processes.

Perhaps more significant for the origin of new functional paralogs and synteny evolution is the discovery of single-gene transposition in plant genomes. Strikingly, Freeling and colleagues (2008) found evidence that 25–75% of all genes in *Arabidopsis* have moved location by single-gene transposition since divergence from *Carica*. Recent analyses in maize (Woodhouse et al. 2010) have uncovered evidence that flanking repeats are associated with many of these transposed genes as well as membership in tandem arrays, suggesting that intrachromosomal recombination between tandem arrays may be responsible for the duplication and movement of these sequences. Evidence for a similar process has been found among other plant lineages, particularly the nucleotide-binding site (NBS)-Leucine-rich repeat (LRR) gene family in *Medicago*. Composed of 400–500 members, the NBS-LRR gene family in *Medicago* appears to have expanded through a combination of tandem duplication and transduplication (Ameline-Torregrosa et al. 2008). Additional high quality genome sequences, especially from relatively closely related plants where mutation and turnover has not completely erased the genomic footprints of duplication mechanisms, are needed to further resolve the contribution of various molecular processes and ultimately evolutionary forces that drive segmental duplications in plant genomes.

11.2.2 Large-Scale Duplications: Polyploidy

Perhaps one of the largest changes ushered in by analyses of plant genomes is a paradigm shift on the importance of polyploidy, or whole genome duplication, in the evolution of plant genome organization and diversity. Prior to genomic analyses of the *Arabidopsis* genome that uncovered evidence of ancient whole genome duplications (Vision et al. 2000), some researchers had come to regard polyploid species as nothing more than evolutionary "dead-ends" (e.g., Wagner 1970). Despite accounting for an abundance of extant plant diversity—nearly 35% of vascular plants are recent polyploids (Wood et al. 2009)—some botanists surmised that polyploidy had, for various reasons, not replaced diploidy as the predominant genetic system and thus polyploids must largely go extinct (Wagner 1970; Stebbins 1971). Although analyses of recent polyploids show that most polyploid species go extinct at much higher rates than related diploids (Mayrose et al. 2011), a fraction of the polyploid species formed over the history of plant evolution have survived and their legacies are evident in plant genomes. Analyses of the *Arabidopsis* genome revealed that a plant long-considered a model diploid species was in fact an ancient polyploid that had experienced at least three rounds of paleopolyploidy that were obscured by gene fractionation and genetic diploidization over millions of years (Vision et al. 2000; Blanc et al. 2003; Bowers et al. 2003). Additional analyses of plant genomic data have uncovered evidence of widespread paleopolyploidy across the plant phylogeny (Blanc and Wolfe 2004a; Pfeil et al. 2005; Cui et al. 2006; Barker et al. 2008, 2009; Tang et al. 2008; Shi et al. 2010), including duplications before the origin of seed plants and angiosperms (Jiao et al. 2011). Far from being an evolutionary dead-end, polyploidy has occurred repeatedly throughout the evolutionary history of plants.

Unlike ongoing small-scale segmental duplications that often produce dead-on-arrival genes or gene fragments, polyploidy punctually creates functional copies of the entire genome. This scale of duplication means that although polyploidies occur at a much lower rate than for example, tandem duplications, they still have a large impact on gene content. The fraction of genes in contemporary "diploid" genomes directly derived from paleopolyploidy varies, but is generally substantial. Near the low end of the range for plants is *Arabidopsis thaliana* with nearly 25% of its genes derived from paleopolyploidy (Blanc et al. 2003). Other plants generally have higher fractions of genes retained from ancient polyploidy: 50% in rice (Wang et al. 2005), 67% in soybean (Schmutz et al. 2010), and ~30% in *Populus* (Tuskan et al. 2006). The time since polyploidy and the number of duplications does not appear to entirely explain the different contributions of polyploidy to plant genomes. For example, *Arabidopsis* genome has been multiplied at least 12 times (Tang et al. 2008) yet contains fewer genes and chromosomes than *Vitis* (Jaillon et al. 2007) whose genome has been multiplied only six times (Tang et al. 2008). Considering that *Arabidopsis* has also experienced two rounds of whole genome duplications since its divergence from *Vitis* (Tang et al. 2008; Barker et al. 2009), it is indicative that variation in rates of gene fractionation and diploidization are just as important as new duplications for shaping the content of plant genomes.

It is important to note that although polyploidy has directly produced a significant fraction of the genes present in many plant genomes, the current complement of genes does not appear to be entirely dependent on polyploidy. A single sequenced plant genome, *Selaginella moellendorffii*, demonstrates no evidence of paleopolyploidy (Banks et al. 2011). However, its genome contains 22,285 annotated genes, nearly 85% of the annotated gene count in *Arabidopsis thaliana* (Haas et al. 2005). *Selaginella's* relatively small genome size and low gene count—the lowest among sequenced land plant genomes—is probably partially attributable to its lack of paleopolyploidy. Although polyploidy has clearly been a predominant force in creating new genes in many plant genomes, the fact that the *Selaginella* genome contains nearly as many genes as an angiosperm whose genome has been multiplied at least 12 times underscores that polyploid history does not, by itself, predict the current complement of genes in plant genomes.

11.3 Differential Fates of Duplicated Genes in Plant Genomes

11.3.1 Introduction and Nonfunctionalization

Newly duplicated genes face several possible fates, but most frequently one member of a newly formed gene pair is lost (Walsh 1995). This is because the rate of deleterious mutation is much higher than that of beneficial mutation (Kimura 1983), and if duplicated genes act redundantly one copy may eventually accumulate deleterious mutations and undergo pseudogenization. Computational simulation supports the view that the rate of gene loss after duplication is at least an order of magnitude higher than gene divergence (e.g., Ohta 1987; Walsh 1995). One need only look at the steepness of the Ks curve for duplicate genes in any eukaryotic genome to see that there are many young duplicates but that relatively few persist. Although the reversion to single copy after gene duplication is expected, there are still numerous duplicated genes retained in all eukaryotic genomes studied. Several different models have been proposed that attempt to explain why and how duplicates are retained (Fig. 11.2). Here we attempt to only provide a more detailed introduction for three predominant models: (1) Neofunctionalization;

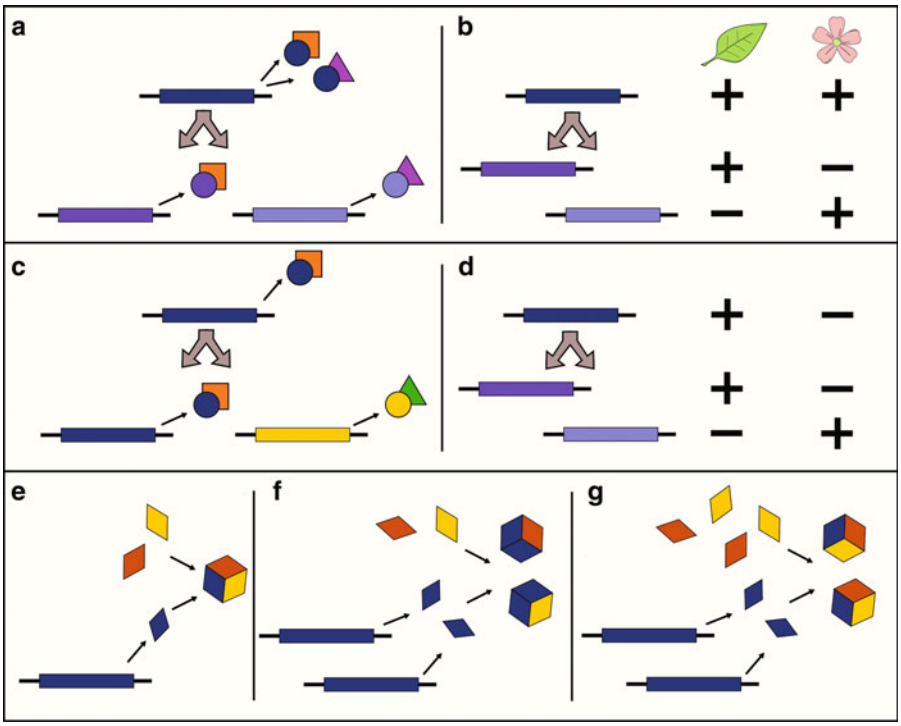

Fig. 11.2 Alternative fates of duplicated genes in plant genomes. Panels (**a**–**d**) *upper* gene model depicts preduplication ancestor, *below* are the duplicates. (**a**) Protein subfunctionalization. (**b**) Expression subfunctionalization. (**c**) Protein neofunctionalization. (**d**) Expression neofunctionalization. Panels (**e**–**g**) portray dosage-sensitive genes and macromolecular complexes produced from their products. (**e**) A dosage sensitive gene contributing a component to a macromolecular product. (**f**) Small scale duplication, gene dosage is unbalanced. (**g**) Large scale duplication resulting in duplication of all genes in the pathway and balanced gene dosage

(2) Subfunctionalization; and (3) Gene dosage balance. More complete discussion of all the proposed routes to duplicate gene retention can be found in recent reviews, e.g., Sémon and Wolfe (2007), Conant and Wolfe (2008), Hahn (2009), and Innan and Kondrashov (2010).

11.3.2 Neofunctionalization

Long hypothesized to be an important source of new variation (Harland 1936; Stephens 1951; Ohno 1970), the neofunctionalization model anticipates the gain of a new function by one gene of a duplicated pair. In this model, duplicated genes are expected to have redundant function so one copy is free to accumulate mutations without affecting the gene's ancestral role. Either rapidly or over a long period of time, this copy may gain a new function. As discussed below with various examples, this process may happen by completely neutral processes (Dykhuizen–Hartl effect), as an adaptive process (positive Darwinian selection), or by some combination of the two.

11.3.2.1 Dykhuizen–Hartl Effect (Neutral Process)

In this scenario, a new function arises as a by-product of the accumulation of chance mutations in the cis-regulatory or protein coding regions. Critically, these neutral mutations become prevalent in the population by genetic drift, a dominant process in eukaryotes with small population sizes. Although initially increased in frequency by non-adaptive mechanisms these mutations might later become adaptive, especially if the ecological or biochemical environment changes, the so called Dykhuizen–Hartl effect (Kimura 1983). Thus, there may be no need for positive Darwinian selection to drive the acquisition of a new function. Liu and Adams (2010) provide a potential example that shows neofunctionalization of a duplicated gene by neutral process. They studied a pair of duplicated genes, *Short Suspensor* (*SSP*) and *Brassinosteroid Kinase 1* (*BSK1*), derived from the most recent Brassicaceae-wide paleopolyploidization, the *At*-α duplication (Tang et al. 2008; Barker et al. 2009). After gene duplication, *SSP* underwent both expression regulatory and protein sequence changes, acquiring a pollen-specific expression pattern and the loss of a protein kinase domain. Changes to both *cis*-regulatory and protein coding regions allowed *SSP* to gain a new function, transitioning from being a component of the brassinosteroid signaling pathway to a paternal regulator of early embryogenesis. Although *SSP* experienced accelerated amino acid changes, a codon-based statistical method for testing positive selection supported a relaxation of purifying selection instead of positive Darwinian selection, suggesting that the neofunctionalization of *SSP* may be the result of neutral processes (Liu and Adams 2010).

11.3.2.2 Positive Darwinian Selection (Adaptive Process)

The second scenario is neofunctionalization by an adaptive process. Positive Darwinian selection may play an important

role in promoting the fixation of beneficial mutations in one member of a new duplicate pair, facilitating neofunctionalization following duplication. Under this model, positive selection drives one copy to gain a new function and thus preserve it—subsequent deleterious mutations to this copy would carry a fitness cost. Several examples of neofunctionalization have been documented in plants. One example of neofunctionalization by positive selection is *MEDEA*, which is a gene duplicated by the most ancient whole genome duplication in Brassicaceae. *MEDEA*, a *SET*-domain *Polycomp* group protein, is a paternally imprinted gene that regulates the development of endosperm in *Arabidopsis*. After duplication, *MEDEA*'s duplicated partner retained the ancestral function whereas *MEDEA* underwent positive Darwinian selection due to parental conflict (Spillane et al. 2007), with a larger magnitude of positive selection in outcrossing lineages in comparison to selfing lineages (Spillane et al. 2007; Miyake et al. 2009). Although it is controversial whether positive selection or balancing selection act on the new function of *MEDEA* (Kawabe et al. 2007), a recent study suggested that both types of selection contributed to the gain of new function (Miyake et al. 2009).

Another example of neofunctionalization driven by positive selection is a pair of functionally redundant cytochrome P450 genes (*CYP98A8* and *CYP98A9*), duplicated by retroposition in Brassicaceae (Matsuno et al. 2009). Their ancestral gene, *CYP98A3*, is involved in the formation of lignin monomer. The ancestral gene is expressed in many organ types but not in pollen, whereas *CYP98A8* and *CYP98A9* are highly expressed in pollen. Their new function has been showed to be involved in a novel phenolic pathway for pollen development. Codon-based methods found that several codons in these two genes have undergone repeated amino acid changes, suggesting that positive Darwinian selection has contributed to the acquisition of new function in *CYP98A8* and *CYP98A9*. Similarly, other cytochrome P450 genes (CYP79s and CYP83s), duplicates from ancient WGDs, have evolved distinct functions in novel biochemical defense pathways (Schranz et al. 2011). However, whether the evolution of these pathways was an adaptive process is still unclear.

11.3.3 Subfunctionalization

An alternative to the long-popular neofunctionalization model is subfuntionalization, which explains duplicate gene retention by the partitioning of multiple ancestral functions or expression patterns among the new paralogs (Hughes 1994; Force et al. 1999). Assuming that a gene has multiple functions or expression patterns before duplication, subfunctionalization would predict that these ancestral functions are simply partitioned among the new paralogs. It should be noted that expression subfunctionalization and protein subfunctionalization will result in two different outcomes. Duplicated genes may share the same function for the subfunctionalization of expression patterns, but both are needed to maintain proper spatial or temporal expression—the loss of one copy will be detrimental to fitness because of mis-expression or loss of expression. For protein subfunctionalization, the ancestral gene must have multiple functions that can be partitioned among the new paralogs. Thus, the loss of a protein subfunctionalized paralog will be deleterious because retention of both copies is essential to fully complement the ancestral function. Such a dichotomous scenario, however, oversimplifies the complexity of possible combinations between expression and protein subfunctionalization and it is possible that the retention of some duplicated genes might be driven by both expression and protein subfunctionalization.

Based on population genetic simulation, subfunctionalization may initially be a neutral stage of duplicate gene evolution, as each copy accumulates mutations that may be reciprocally deleterious without interrupting the total function of their ancestral state. Thus, it has been argued that subfunctionalization is likely more important than neofunctionalization, especially in organisms with small effective population sizes, because it is easier to lose an existing function by mutation than gain a new one (Lynch and Force 2000). As we discuss in more detail below, this does not necessarily mean that positive selection will not act on subfunctionalized gene duplicates for long-term preservation. Subfunctionalization may be further classified into two different models: (1) duplication, degeneration, and complementation (DDC) model and (2) escape from adaptive conflict (EAC) model.

11.3.3.1 DDC Model (Neutral Process)

When Force et al. (1999) initially proposed the DDC model they focused on regulatory aspects of duplicate gene evolution. According to this model, both copies will neutrally accumulate deleterious mutations on their cis-regulatory regions. After sufficient mutation accumulation significantly impairs function, each copy will retain only a fraction of the ancestral phenotype (i.e., subfunction) and complement each other to cover the full spectrum of their ancestral expression pattern. Thus, this kind of subfunctionalization has been termed as duplication-degeneration-complement (DDC) model. It has been shown that the probability of DDC subfunctionalization is much higher than that of regulatory neofunctionalization, especially in organisms with small effective population sizes where positive selection is less efficient and genetic drift is more predominant (Force et al. 1999). A nice example of the DDC model in plants was illustrated by Force et al. (1999). A pair of MADS-box transcription factor genes (*ZAG1* and *ZMM2*) in maize,

which originated from an allopolyploidization event, showed reciprocal expression pattern where *ZAG1* is highly expressed throughout carpel development but weakly expressed in stamen, and *ZMM2* is highly expressed in stamen (Mena et al. 1996). In comparison to their putative ancestral expression pattern from a single orthologous gene in *Arabidopsis (AGAMOUS)* and *Antirrhinum (PLENA)* that are highly expressed in both carpels and stamen (Yanofsky et al. 1990; Coen and Meyerowitz 1991; Bradley et al. 1993), it is reasonable to infer that *ZAG1* and *ZMM2* have subfunctionalized (i.e. DDC) since their duplication. Another example consistent with the DDC model is an organ-specific reciprocal gene silencing pattern for a pair of homoeologous alcohol dehydrogenase genes (*AdhA*) in allopolyploid cotton (Adams et al. 2003). In this case, one copy is only expressed in petals and stamen whereas the other copy is only expressed in carpels, indicative of regulatory subfunctionalization across different organ types. Because the polyploidization event in cotton has been inferred to occur about 1–2 million years ago (Cronn et al. 2002; Senchina et al. 2003) and polyploid plants likely have small effective population sizes at the onset of polyploidization, the *AdhA* example supports the perspective that the DDC model is important for the retention of duplicated genes in organisms where genetic drift predominates their population genetic landscape.

11.3.3.2 EAC Model (Adaptive Process)

Under the Escape from Adaptive Conflict (EAC) model, a new function arises in the ancestral gene but it reduces the performance of the original function. This creates an adaptive conflict where improving either function comes at a cost to the other. One solution to resolve such adaptive conflicts is by gene duplication. After duplication, one copy is free to improve the new function whereas the other copy may improve the original function. Discriminating subfunctionalization by EAC and neofunctionalization by adaptive process relies on these improvements because both processes will leave signatures of positive Darwinian selection. However, both copies will experience adaptive changes under the EAC model whereas only one copy will show adaptive change if neofunctionalization occurs. Further, the ancestral function will be improved if EAC took place but not under neofunctionalization. Des Marais and Rausher (2008) report a nice example of the EAC model in tandemly duplicates of the anthocyanin biosynthesis pathway gene dihydroflavonol-4-reductase (*DFR*) in morning glories (*Ipomea*). The major function of *DFR* is the regulation of flower color by via the anthocyanin biosynthesis pathway. After the first gene duplication event, one copy (*DFR-A*/*C* clade) underwent repeated positive Darwinian selection based on codon-based models of sequence evolution while the copy (*DFR-B* clade) showed an improvement of its ancestral function by increasing its enzyme activity on dihydrokaempferol (DHK), dihydroquercetin (DHQ) and dihydromyricetin (DHM). In addition, the copy in *DFR-A*/*C* clade has been demonstrated to lose its ancestral function, suggesting that a new function has been acquired after gene duplication although this new function remains uncharacterized.

11.3.4 Frequency of Subfunctionalization and Neofunctionalization on a Genome-Wide Scale

A number of examples demonstrate that both neofunctionalization and subfunctionalization contribute to the evolution of new genes in plant genomes. However, the relative contribution of these two processes to duplicate gene retention is a major question in the evolution of plant genome content. Large-scale genomic studies, as outlined below, are being used to investigate expression pattern and protein sequence divergence to make inferences about the importance of sub- and neofunctionalization.

11.3.4.1 Expression Pattern Divergence

Genome-wide studies of expression divergence between duplicated genes have largely focused on rice, *Arabidopsis*, and cotton. Blanc and Wolfe (2004b) and Haberer et al. (2004) investigated the expression divergence between duplicated genes in *Arabidopsis*. Overall, they found that the majority of duplicated genes have significantly divergent expression patterns. For example, 57% of *At*-α whole genome duplicated genes and 73% of older *At*-β or *At*-γ whole genome duplicated genes have diverged in their expression pattern from 62 different developmental stages and organ types, indicating that many paralogs may functionally diversify after duplication even though protein divergence between most duplicated genes is small (Blanc and Wolfe 2004a, b; Haberer et al. 2004). In rice, 88% of paralogs had divergent expression (Throude et al. 2008). The correlation of expression divergence and protein divergence is indeed observed in later studies where Casneuf et al. (2006) and Ganko et al. (2007) found that older paralogs showed more functional divergence than younger gene duplicates. A similar trend was also found in rice duplicated genes (Li et al. 2009). Further, studies by Casneuf et al. (2006) and Ganko et al. (2007) demonstrated that expression pattern in small-scale duplications (i.e., tandem duplication or dispersed duplication) are less correlated than those in large-scale duplications (i.e., chromosomal or whole genome duplication), suggesting that small-scale duplications potentially have a higher chance to be neofunctionalized than large-scale duplications. Although these studies demonstrate that paralogs frequently have different expression patterns, they are unable to infer regulatory subfunctionalization or

neofunctionalization due to the lack of information about their ancestral expression pattern.

The first attempt to ascertain the relative contribution of regulatory sub- or neofunctionalization for the retention of duplicated genes in plants on a genome-wide scale was recently conducted by Zou and colleagues (2009). They used the stress response of paralogs to gain a better understanding of the relative contributions of regulatory sub- and neofunctionalization in duplicate evolution. Using a maximum likelihood probabilistic framework with gene family phylogenies, they employed a Bayesian method to reconstruct the putative most recent common ancestral expression pattern for every duplicated gene pair in *Arabidopsis*. From their analysis, ~61% duplicated genes retained ancestral responsiveness to stress, ~30% showed the loss of ancestral responsiveness to stress, ~6% experienced the gain of a new response to stress, and only ~2% experienced a functional switch that changed their response to the opposite of their ancestral state. For the ~30% of paralogs where the ancestral expression pattern was lost, the ancestral functions were partitioned between gene duplicates in most cases, suggesting that subfunctionalization plays an important role in paralog retention in response to stresses. Their analyses also found that older duplicates tended to be more neofunctionalized than younger paralogs, which were more often subfunctionalized. Although Zou et al. (2009) found that subfunctionalization is more frequent than neofunctionalization, it should be noted that they only analyzed the gene expression pattern under different stress conditions. It is reasonable to assume that the relative contribution of subfunctionalization and neofunctionalization will differ when someone survey different dataset (e.g., developmental stages, cell types, organ types) or more conditions (e.g., temporal and spatial). For example, Duarte et al. (2006) found that 85% of 280 *Arabidopsis* duplicated pairs showed significant gene by organ effect across six different organs based on an analysis of variance approach, indicative of potentially regulatory subfunctionalization or regulatory neofunctionalization.

How fast can the regulatory sub-/neofunctionalization arise after gene duplication? To answer this question, Adams et al. (2003) and Chaudhary et al. (2009) surveyed, in total, 103 duplicated genes in various tissues/organ types in recently formed polyploid cotton, *Gossypium hirsutum*. Interestingly, eight cases of regulatory subfunctionalization (~8%) and 15 cases of probable regulatory neofunctionalization (~15%) were observed in their analyses. Their studies provide striking examples that regulatory subfunctionalization and neofunctionalization can quickly arise soon after genome duplication, suggesting that both mechanisms are important for the retention of duplicated genes over evolution.

11.3.4.2 Protein Sequence Divergence

Neofunctionalization predicts that gene duplicates will evolve asymmetrically. In other words, one copy will experience relaxed purifying selection, accumulate more nonsynonymous substitution or positive selected sites, whereas the other copy will experience strong purifying selection and accumulate mutations without evidence of significant amino acid change or positive selection. Based on these assumptions, Blanc and Wolfe (2004a, b) surveyed the asymmetric rate of sequence evolution between duplicated genes in *Arabidopsis* to assess the relative contribution of neofunctionalization in the retention of duplicated genes in plants. Nearly 21% of paralogs derived from the *At*-α whole genome duplication showed asymmetric sequence rate evolution, indicative of potential neofunctionalization. Their observation suggests that neofunctionalization may play an important role in paralog retention in plants. However, it should be noted that using asymmetric sequence rate analysis is not a perfect test for neofunctionalization. As mentioned previously, asymmetric sequence rate evolution is also possible under the EAC model and it is difficult to use difference in paralog rate evolution as a method to distinguish between subfunctionalization and neofunctionalization.

To resolve this issue, better knowledge of the ancestral protein function or structure, pre-duplication, will be important. However, characterization of gene function is a tedious and daunting task and plant biologists often must apply a case-by-case approach. It is therefore difficult to apply a rigorous, genome-wide analysis in this regard. Analyses of protein structure may be high-throughput, and those related to protein subcellular relocalization (PSR) offer a promising approach to evaluate post-duplication fates. PSR is often associated with the 5′ N-terminal signal peptide (detailed review in Byun-McKay and Geeta 2007). By comparing this region with other closely related species, it is possible to infer if the change of subcellular localization results from subfunctionalization or neofunctionalization. An unpublished analysis of differential subcellular localization found that the N-terminal peptide in *Arabidopsis* duplicated genes evolved rapidly relative to the rest of the protein region, implying that PSR might greatly contribute to functional diversification between duplicated genes in plants (Byun-McKay et al. 2009). So far, studies that aim to understand the relative importance of subfunctionalization and neofunctionalization after gene duplication at the protein level on a genome-wide scale remain rare. In addition to rice and *Arabidopsis*, more genomic resources from closely related plant species will help us to understand the relative contribution of protein subfunctionalization and neofunctionalization for the retention of duplicated genes in plants.

11.3.5 Gene Dosage Balance

The concept of dosage balance arose from studies of chromosome segregation in plants during the early twentieth century (e.g., Blakeslee and Avery 1919; Blakeslee et al. 1920; Blakeslee 1921). Pioneering plant geneticists found that unbalanced chromosome numbers such as trisomy or monosomy often caused significant phenotypic changes. Some of these phenotypic changes were strongly deleterious and severely impacted plant growth. Based on these observations, plant geneticists proposed that imbalance in the products of genetic materials carried by the chromosomes were responsible for these phenotypic changes (Blakeslee 1921). Support for this hypothesis was found in examples of chromosomal changes that restored proper dosage. For example, a secondary isochromosome (fusion of two sister chromatids) in a trisomic plant would alleviate the phenotypic changes of a pure trisomic line (Belling and Blakeslee 1924). In this case, the secondary isochromosome yields four identical chromosomal arms, which consequently produced proper gene dosage balance and ameliorate deleterious phenotypic changes. These concepts have been updated recently by researchers utilizing molecular genetic approaches to study the effects of gene copy number, or dosage, on phenotype and gene duplication in yeast (Veitia 2002, 2003; Papp et al. 2003), fruitfly, and plants (Birchler et al. 2001). Significantly, this recent work has found that the products of genes involved in regulatory roles (e.g., transcriptional factors and signal transductions) appear to require maintenance of their stoichiometric relationships, or dosage balance, likely because they participate in the formation of macromolecular complexes (see review by Birchler and Veitia 2007 and references therein). These concepts have been synthesized into the gene dosage balance hypothesis, whereby genes involved in more interaction, so called "connected" genes, are more dosage-sensitive (see introduction by Freeling and Thomas 2006).

How does gene dosage influence the retention of duplicated genes in plants? After gene duplication, gene dosage balance is commonly invoked to explain that either gene loss or retention is required to maintain the proper amount of protein product stoichiometry, or dosage, in genomes (e.g., Birchler et al. 2001; Veitia 2002, 2003; Papp et al. 2003). Based on this model, duplicated genes that have large pleotropic effects (or more interactions with other proteins) are expected to be retained more often after a large scale duplication event (i.e., whole genome duplication) than a small scale duplication event (i.e., tandem duplication, dispersed duplication, and retroposition). This is because other interacting partners will also be duplicated in large-scale events and retention of all genes is necessary to maintain stoichiometry, whereas not all interacting loci will be duplicated in a small-scale event and loss of the new paralog is required for proper dosage (see review by Freeling 2009). Specifically, the dosage balance hypothesis predicts that genes such as transcription factors or involved in signal transduction are expected to be retained after whole genome duplication such that entire genetic network still has proper amount of upstream regulator to regulate the doubled downstream pathway. If such highly connected genes are created by small scale events they would disrupt the non-duplicated downstream elements. In contrast, genes involved in the terminals of genetic networks are not prevented from being retained following small-scale duplications because additional copies of these genes is less likely to cause significant imbalance on the genetic network.

As expected, genes involved in transcriptional factors and signaling transduction often are overrepresented in duplicated genes from whole genome duplication while genes involved in abiotic or biotic stresses are overrepresented in those from tandem duplication in *Arabidopsis*, rice, and kiwifruit (Blanc and Wolfe 2004a, b; Tian et al. 2005; Shi et al. 2010). Further evidence supporting gene dosage balance and the retention of duplicated genes was recently reported by Liang et al. (2008). They investigated the relationship between protein under-wrapping (i.e., a molecular quantifier of the reliance of the protein on binding partnerships to maintain its protein structure stability or, simply the degrees of interaction with other molecules) and gene duplicability across various organisms, including *E. coli*, yeast, nematodes, fruit fly, mammals, and plants. A strong negative correlation between protein under-wrapping and gene duplicability was found, suggesting that dosage-sensitive genes are unlikely to be duplicated by small-scale duplication events. Their analysis supports the argument that proteins with highly interacted partners are more sensitive to dosage imbalance and be less like to be retained. It should be noted that not all analyses of ancient polyploidy in plants have found results consistent with dosage balance hypothesis. In the Compositae, Barker et al. (2008) found the opposite pattern predicted by the dosage balance hypothesis; signaling genes and transcription factors were significantly lost across whole genome duplications in multiple species, whereas metabolic and structural genes were strongly retained in duplicate (Fig. 11.3). It may be that the dosage-sensitivity of gene families varies across the plant phylogeny, and further study in model systems outside of the Brassicales would be valuable to better understand the evolution of plant genomes in these disparate lineages.

One intriguing possibility is that the shrinkage or expansion of plant gene families across the phylogeny may be largely attributed to the degree of dosage sensitivity. Cannon et al. (2004) investigated the relative contribution of large-scale and tandem gene duplication on the evolution of 50 different large gene families in *Arabidopsis*. As predicted by gene dosage balance hypothesis, gene families that function

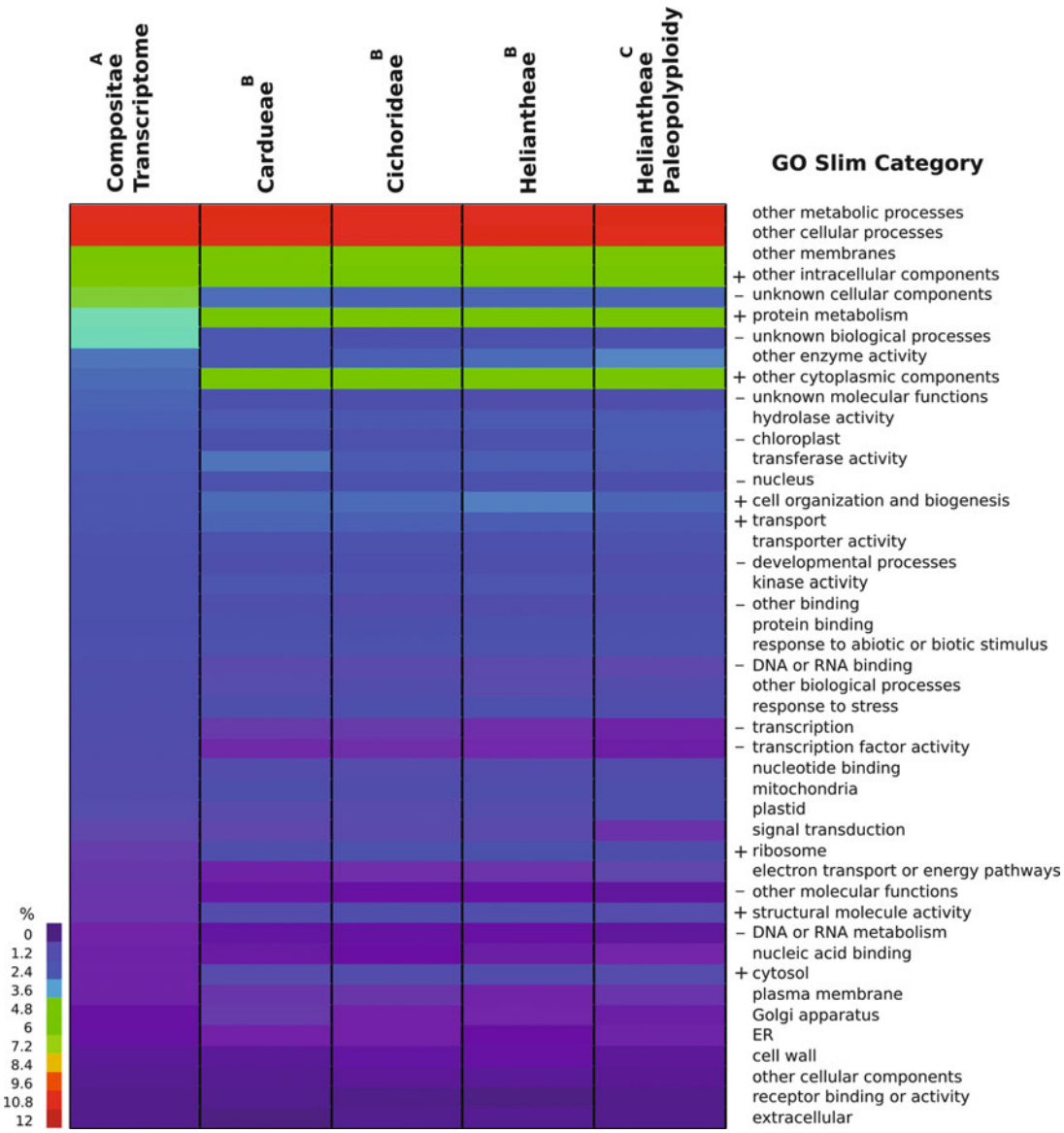

Fig. 11.3 Cell plot of gene ontology frequencies demonstrating biased gene retention following ancient polyploidy in the Compositae (Barker et al. 2008)

as transcription factor (e.g., MYB transcription factor) and signaling molecules (e.g., GTP binding protein and calmodulin) contained a higher proportion of genes derived from large-scale duplications, whereas gene families that function as pathogen defenses (e.g., major latex protein and chlorophyll a-b binding) were composed of a high proportion of tandemly duplicated genes. Such an observation further supports the importance of gene dosage balance on the retention of duplicated genes, whereby more "connected" genes (i.e., upstream regulator of genetic network such as transcription factor and signaling) would be likely retained after large scale duplication and less "connected" genes (i.e., terminal nodes of genetic network such as genes in response to biotic or abiotic stresses) should be likely retained by small scale duplication. This reciprocal pattern of retention (i.e. WGD versus tandem duplicates) is supported by the distribution of duplicates across the entire *Arabidopsis* metabolic network (Bekaert et al. 2011). The retained duplicates from ancient WGDs were significantly more clustered (i.e. "connected") and involved in high flux reactions across the entire network compared to tandem duplicates. Similarly, Freeling (2009) observed a reciprocal pattern in gene retention following WGDs compared to tandem duplications across various GO terms.

Under the gene dosage balance hypothesis, what will be the evolutionary driving force operating on the retention of duplicated genes? Veitia (2005) illustrated that mutations for genes encoding multidomain proteins can produce dominant

negative phenotypes, thereby inactivating the macromolecular complexes. For example, within the context of a dimer AA, one mutant allele in a tetraploid will result in only 56% of the AA dimers being active. Assuming that the decrease of active macromolecular complexes reduces fitness, such dominant negative effects may serve as strong purifying selection for the retention and maintenance of duplicated dosage-sensitive genes (see detail illustrations and description in Veitia 2005). In contrast, the retention of numerous copies of dosage-insensitive genes could be driven by positive Darwinian selection. For example, Hanikenne et al. (2008) provided a nice example about the expansion of small-scale duplicated genes in related to abiotic stress where the tolerance of high heavy metal concentration in *Arabidopsis halleri* was due to the triplication of three tandemly arrayed heavy metal aptase 4 (*HMA4*) genes. The amplification of copy number for HMA4 allows *A. halleri* to colonize heavy-metal polluted soil, suggesting that the retention of small-scale duplicated genes were largely driven by positive selection. In addition, the authors showed that enhanced expression level of HMA4 by changes of cis-element region also contribute to the tolerance of heavy metal in *A. halleri*. Thus, the increase of protein product itself, rather than maintenance of stoichiometry, may be beneficial for fitness in some circumstances.

Which model best explains the patterns of retention and loss of duplicated genes in plant genomes? Although all models likely act in plant genomes, gene dosage balance has become favored by many researchers because it makes explicit predictions about retention and loss. Neofunctionalization and subfunctionalization models explain how and why duplicate genes may be retained, but they are unable to generate predictions with regards to which genes will be retained after different types of duplication. Only gene dosage balance model can predict the reciprocal gene function retention pattern between whole genome duplicates and tandem duplicates in several plants such as *Arabidopsis*, rice, and kiwifruit (Freeling 2009; Edger and Pires 2009; Shi et al. 2010). However, these models are not mutually exclusive. For example, it is likely that the retention of dosage-insensitive gene duplicates from small-scale duplications, which are essentially dosage-neutral, may often be best explained by the sub- and neofunctionalization models. Thus, it is likely that all of these models contribute to the content of plant genomes with their efficacies varying across gene families and duplication type.

11.4 Duplication, Turnover, and Re-organization of Plant Genomes

All plant species have at least undergone one round of whole genome duplication over their evolutionary history (Blanc and Wolfe 2004a, b; Pfeil et al. 2005; Cui et al. 2006; Barker et al. 2008; Tang et al. 2008; Shi et al. 2010; Jiao et al. 2011). However, most of these species demonstrate diploid chromosomal behavior and disomic inheritance restored through a process termed diploidization (Wolfe 2001). Diploidization is a dynamic process of fractionation, shuffling, and divergence of duplicated portions of a genome. The scale of events that contribute to diploidization range from loss of single genes to entire chromosomes. Various aspects of the process have been studied at time scales ranging from early generation polyploids to fully diploidized paleopolyploids. Diploidization results in the large scale re-organization of plant genomes and parallels the process of gene duplication, gene loss and functional divergence (Fig. 11.4). In this section, we present recent work on the effect of whole genome duplication on the dynamics of genomic reorganization from early generation polyploids to fully diploidized paleopolyploids in plants. These examples

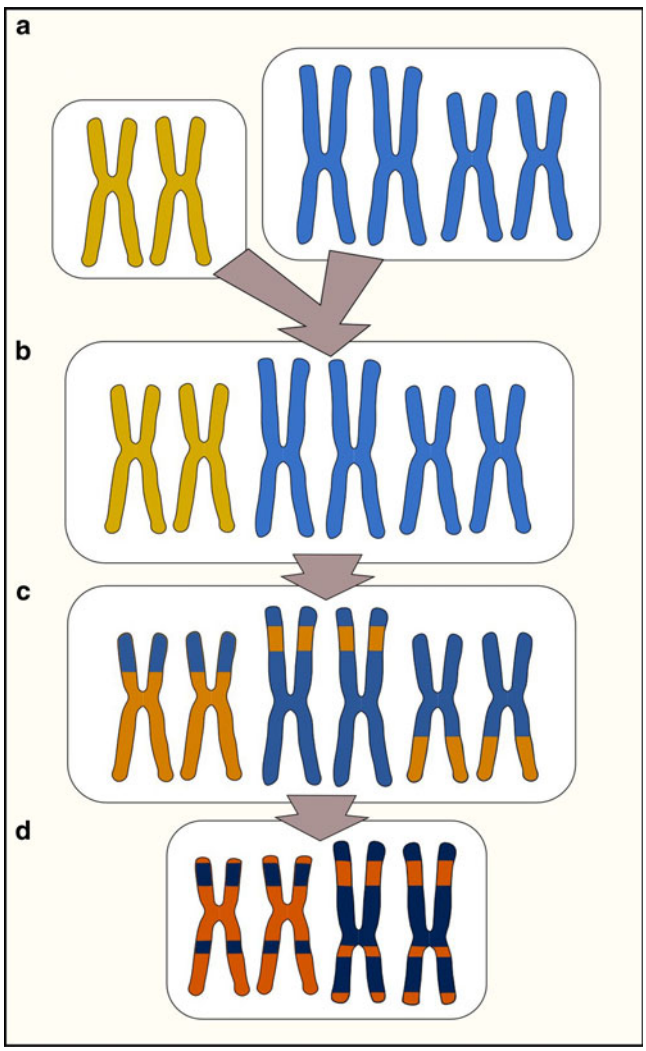

Fig. 11.4 Schematic of progressive stages of genomic diploidization following an ancient whole genome duplication depicting homoeologous pairing and chromosomal loss

will provide some insight on our current understanding of turnover and genome evolution in plants.

11.4.1 Re-organization of Polyploid Plant Genomes

Diploidization may begin as early as the first generation following genome duplication. Chromosomal rearrangements have been observed in many early generation polyploids. For example, in 50 synthetic allopolyploid lines of *Brassica napus* genomic changes were present but rare in the first generation (Lukens et al. 2006) and by the fifth generation every line had evidence for recombination between homologous chromosomes. These recombination events resulted in the loss of alleles from one of the parents and a doubling of the other. A recent analysis of synthetic *B. napa* allopolyploid lines found variation in chromosome number as early as the first generation (Xiong et al. 2011). Further, Xiong and colleagues (2011) discovered that although there was extensive loss and shuffling of chromosomes, the total number was maintained close to the tetraploid count with apparent compensatory replacement of homoeologous sets, putatively to maintain dosage balance. The rate of these early generation changes appears to vary significantly across lineages, even within the genome of a species. Allohaploid lines from a diverse panel of *B. napus* were found to have markedly different rates of homoeologous recombination and suggest different levels of crossover suppression between homoeologs (Cifuentes et al. 2010). Similar early generation changes have also been found in natural polyploids. A striking example is the *Tragopogon* allopolyploids, which formed less than 80 years ago (Soltis et al. 2004), but which have experienced a variety of genomic changes following polyploidy, including chromosomal rearrangements and loss (Lim et al. 2008). However, not all polyploids demonstrate significant changes in genome organization. The genomes of nearly cotton allopolyploids demonstrate very little genomic changes (Liu et al. 2001). It is not clear why some plants, such as cotton, do not experience rapid changes in chromosomal rearrangement after polyploidization. At larger time scales, the rate of synteny evolution is clearly variable across plants (Jaillon et al. 2007), and it may be a similar or the same process at work here. Future work associating the rates of turnover and synteny evolution with life history traits, generation time, and effective population sizes may be fruitful avenues for advancing our understanding of these patterns of plant genome evolution.

The process of chromosomal rearrangements and sequence divergence is continuous following polyploid formation. Allopolyploid tobacco species with similar parentage but a range of times since formation have been observed to follow this expectation. In a recent study, young polyploids were found to have fewer rearrangements and less sequence divergence than older ones and the most ancient allopolyploid of the group (~5 million years old) experienced too much sequence divergence for the technique employed (Genomic In Situ Hybridization) to identify the parental alleles within its genome (Lim et al. 2007). Australian relatives of *Arabidopsis* experienced a lineage-specific whole genome duplication event since the well documented *At*-α WGD, 6–9 million years ago. Strikingly, the three species studied have a relatively low chromosome counts of $n = 4$, 5, and 6 with the genetic behavior of diploids. Reconstruction of their polyploid ancestor suggests there would have been 16 chromosomes in their ancestral genomes. Mandáková and colleagues (2010) found that most segments of their genomes still exist in duplicate although extensive and independent rearrangement has occurred in these different species. The large amount of chromosome number reduction would have required multiple chromosome loss and fusion events as well as the loss of several centromeres (Mandáková et al. 2010). Most rearrangements in plant genomes probably involve high repeat regions, telomeres and centromeres (Lysak et al. 2006), and in this case it seems that substantial amounts of end-to-end chromosomal fusion and unequal reciprocal translocations were the main mechanisms of chromosomal reduction. Whether or not such cases are outstanding or routine among plant lineages remains to be seen.

An interesting observation made when comparing syntenic blocks of ancient whole genome duplicates is a bias in which of the blocks undergo fractionation or loss of duplicates. The majority of syntenic blocks in *Arabidopsis* remaining from the *At*-α WGD have biased fractionation where clusters of genes were retained within blocks on the same homoeolog, but lost from the other (Thomas et al. 2006). Surprisingly, Thomas and colleagues (2006) found that ~85% of the *Arabidopsis* genome experienced this biased fractionation, most often involving genes in the same genetic network consistent with the dosage-balance hypothesis. A similar study in maize, which had the benefit of having an extant pre-duplication outgroup for comparison, also found evidence of biased fractionation. The authors propose a mechanism of the deletion of homologs by small intrachromosomal recombination (Woodhouse et al. 2010). It is possible that some of these differences existed in the ancestral parental genomes before the genome merger and duplication event; however, research with younger polyploids suggest that the biased gene content in the different blocks is due to biased fractionation following polyploidy. For example, the young (<80 years) allotetraploid *Tragopogon mirus* also exhibits biased loss towards one parent (Koh et al. 2010). Significantly, synthetic F1 crosses between the two parental species demonstrated

only additive patterns of both locus presence and expression. These results suggest that the biased patterns of homoeolog loss observed in natural polyploid populations are the result of later generation changes and are not the immediate result of hybridization. Biased fractionation following WGD has also been documented in protozoan, fungal, and animal lineages (e.g. Paramecium, yeast, teleost; Sankoff et al. 2010). Further studies on patterns of gene expression, homoeolog loss, cytonuclear interactions, and epigenetic phenomena in related allo- and autopolyploids are needed to disentangle the effects of hybridization and polyploidy on biased fractionation in plant genomes.

Conclusions

Duplication and loss are fundamental forces in the evolution of plant genomes. Polyploidy and smaller scale duplications, while operating at different scales and frequencies, have provided the raw material and set the pace for the evolution of plant gene families as well as genome organization. Differential loss and retention of paralogs has led to significantly different gene family content across the vascular plants (Banks et al. 2011) that were likely important for the evolution of the defining phenotypes of many lineages of land plants (Banks et al. 2011; Jiao et al. 2011) and their evolutionary success. The relatively high cycles of duplication and loss experienced by plants have also driven substantial changes in genome organization, yielding more rapid changes among the genomes of related plants compared to animals (Salse et al. 2009). Resolving the relationships among repetitive element content, rearrangements, chromosomal loss and gain, dosage balance, and life history features promises to be an exciting synthesis of plant genome evolution over the next decade. Although much progress has been made in the last decade, the growing genomic data from across plants will refine our knowledge of plant genomes and population and clade-scale genomics will provide the ultimate resources to test the numerous hypotheses proposed by plant geneticists over the past century.

References

Adams KL, Cronn R, Percifield R, Wendel J (2003) Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. Proc Natl Acad Sci U S A 100:4649–4654

Ameline-Torregrosa C, Wang B-B, O'Bleness MS, Deshpande S, Zhu H, Roe B, Young ND, Cannon SB (2008) Identification and characterization of nucleotide-binding site-leucine-rich repeat genes in the model plant *Medicago truncatula*. Plant Physiol 146:5–21

Banks JA et al (2011) The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. Science 332:960–963

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

Barker MS, Vogel H, Schranz ME (2009) Paleopolyploidy in the Brassicales: analyses of the *Cleome* transcriptome elucidate the history of genome duplications in *Arabidopsis* and other Brassicales. Genome Biol Evol 1:391–399

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

Belling J, Blakeslee AF (1924) The configurations and sizes of the chromosomes in trivalents of 25-chromosome *Daturas*. Proc Natl Acad Sci U S A 10:116–120

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

Birchler JA, Bhadra MP, Auger DL (2001) Dosage-dependent gene regulation in multicellular eukaryotes: implications for dosage compensation, aneuploid syndromes, and quantitative traits. Dev Biol 234:275–288

Blakeslee AF (1921) Types of mutations and their possible significance in evolution. Am Nat 5:254–267

Blakeslee AF, Avery BT (1919) Mutations in the Jimson weed. J Hered 10:111–120

Blakeslee AF, Belling J, Farnham ME (1920) Chromosomal duplication and Mendelian phenomena in *Datura* mutants. Science 52:388–390

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

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

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

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

Bradley D, Carpenter R, Sommer H, Hatley N, Coen E (1993) Complementary floral homeotic phenotypes result from opposite orientations of a transposon at the plena locus of *Antirrhinum*. Cell 72:85–95

Byun-McKay SA, Geeta R (2007) Protein subcellular relocalization: a new perspective on the origin of novel genes. Trends Ecol Evol 22:338–344

Byun-McKay SA, Geeta R, Duggan R, Carroll B, McKay SJ (2009) Missing the subcellular target: a mechanism of eukaryotic gene evolution. In: Pontarotti P (ed) Evolutionary biology: concept, modeling, and application. Springer, Berlin/Heidelberg/New York, pp 175–183

Cannon SB, Mitra A, Baumgarten A, Young ND, May G (2004) The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. BMC Plant Biol 4:10

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

Chaudhary B, Flagel L, Stupar RM, Udall JA, Verma N, Springer NM, Wendel J (2009) Reciprocal silencing, transcriptional bias and functional divergence of homoeologs in polyploid cotton (*Gossypium*). Genetics 182:503–517

Cifuentes M, Eber F, Lucas M-O, Lode M, Chèvre A-M, Jenczewski E (2010) Repeated polyploidy drove different levels of crossover

suppression between homoeologous chromosomes in *Brassica napus* allohaploids. Plant Cell 22:2265–2276
Coen ES, Meyerowitz EM (1991) The war of the whorls: genetic interactions controlling flower development. Nature 353:31–37
Conant GC, Wolfe KH (2008) Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet 9:938–950
Cronn RC, Small RL, Haselkorn T, Wendel J (2002) Rapid diversification of the cotton genus *(Gossypium:* Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. Am J Bot 89:707–725
Cui LY, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A, Albert VA, Ma H, dePamphilis CW (2006) Widespread genome duplications throughout the history of flowering plants. Genome Res 16:738–749
Des Marais DL, Rausher MD (2008) Escape from adaptive conflict after duplication in an anthocyanin pathway gene. Nature 454:762–765
Doyle JJ, Flagel LE, Paterson AH, Rapp RA, Soltis DE, Soltis PS, Wendel JF (2008) Evolutionary genetics of genome merger and doubling in plants. Annu Rev Genet 42:443–461
Duarte JM, Cui L, Wall PK, Zhang Q, Zhang X, Leebens-Mack J, Ma H, Altman N, dePamphilis CW (2006) Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of *Arabidopsis*. Mol Biol Evol 23:467–478
Edger PP, Pires JC (2009) Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. Chromosome Res 17:699–717
Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545
Freeling M (2009) Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annu Rev Plant Biol 60:433–453
Freeling M, Thomas BC (2006) Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res 16:805–814
Freeling M, Lyons E, Pederson B, Alam M, Ming R, Lisch D (2008) Many or most genes in *Arabidopsis* transposed after the origin of the order Brassicales. Genome Res 18:1924–1937
Ganko EW, Meyers BC, Vision TJ (2007) Divergence in expression between duplicated genes in *Arabidopsis*. Mol Biol Evol 24:2298–2309
Haas BJ et al (2005) Complete reannotation of the *Arabidopsis* genome: methods, tools, protocols and the final release. BMC Biol 3:7
Haberer G, Hindemitt T, Meyers BC, Mayer KFX (2004) Transcriptional similarities, dissimilarities, and conservation of cis-elements in duplicated genes of *Arabidopsis*. Plant Physiol 136:3009–3022
Hahn MW (2009) Distinguishing among evolutionary models for the maintenance of gene duplicates. J Hered 100:605–617
Haldane JBS (1933) The part played by recurrent mutation in evolution. Am Nat 67:5–9
Hanikenne M, Talke IN, Haydon MJ, Lanz C, Nolte A, Motte P, Kroymann J, Weigel D, Kramer U (2008) Evolution of metal hyperaccumulation required cis-regulatory changes and triplication of HMA4. Nature 453:391–395
Harland SC (1936) The genetical conception of the species. Camb Philos Soc Biol Rev 11:83–112
Hiller LDW et al (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432:695–716
Hoen DR, Park KC, Elrouby N, Yu Z, Mohabir N, Cowan RK, Bureau TE (2006) Transposon-mediated expansion and diversification of a family of *ULP*-like genes. Mol Biol Evol 23:1254–1268
Hughes AL (1994) The evolution of functionally novel proteins after gene duplication. Proc R Soc Lond B Biol Sci 256:119–124
Innan H, Kondrashov F (2010) The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet 11:97–108
Jaillon O et al (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467
Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR (2004) Pack-MULE transposable elements mediate gene evolution in plants. Nature 431:569–573
Jiao Y, Wickett NJ, Ayyampalayam S, Chanderball AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Llang H, Soltis PS, Soltis DE, Clifton SW, Schlarbaum SE, Schuster SC, Ma H, Leebens-Mack J, dePamphilis CW (2011) Ancestral polyploidy in seed plants and angiosperms. Nature 473:97–100
Juretic N, Hoen DR, Huynh ML, Harrison PM, Bureau TE (2005) The evolutionary fate of MULE-mediated duplications of host gene fragments in rice. Genome Res 15:1292–1297
Kawabe A, Fujimoto R, Charlesworth D (2007) High diversity due to balancing selection in the promoter region of the Medea gene in *Arabidopsis lyrata*. Curr Biol 17:1885–1889
Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge
Koh J, Soltis PS, Soltis DE (2010) Homeolog loss and expression changes in natural populations of the recently and repeatedly formed allotetraploid *Tragopogon mirus* (Asteraceae). BMC Genomics 11:97
Leitch IJ, Soltis DE, Soltis PS, Bennett MD (2005) Evolution of DNA amounts across land plants (Embryophyta). Ann Bot 95:207–217
Li Z, Zhang H, Ge S, Gu X, Gao G, Luo J (2009) Expression pattern divergence of duplicated gene in rice. BMC Bioinformatics 10:S8
Liang H, Plazonic KR, Chen J, Li WH, Fernández A (2008) Protein under-wrapping causes dosage sensitivity and decreases gene duplicability. PLoS Genet 4:e11
Lim KY, Kovarik A, Matyasek R, Chase MW, Clarkson JJ, Grandbastien MA, Leitch AR (2007) Sequence of events leading to near-complete genome turnover in allopolyploid *Nicotiana* within five million years. New Phytol 175:756–763
Lim KY, Soltis DE, Soltis PS, Tate J, Matyasek R, Srubarova H, Kovarik A, Pires JC, Xiong Z, Leitch AR (2008) Rapid chromosome evolution in recently formed polyploids in *Tragopogon* (Asteraceae). PLoS One 3:e3353
Liu S-L, Adams KL (2010) Dramatic change in function and expression pattern of a gene duplicated by polyploidy created a paternal effect gene in the Brassicaceae. Mol Biol Evol 27:2817–2828
Liu B, Brubaker CL, Mergeai G, Cronn RC, Wendel JF (2001) Polyploid formation in cotton is not accompanied by rapid genomic changes. Genome 44:321–330
Lukens LN, Pires JC, Leon E, Vogelzang R, Oslach L, Osborn T (2006) Patterns of sequence loss and cytosine methylation within a population of newly resynthesized *Brassica napus*. Plant Physiol 140:336–348
Lynch M (2007) The origins of genome architecture. Sinnauer, Sunderland
Lynch M, Conery JS (2003) The evolutionary demography of duplicate genes. J Struct Funct Genomics 3:35–44
Lynch M, Force AG (2000) The origin of interspecific genomic incompatibility via gene duplication. Am Nat 156:590–605
Lysak MA, Berr A, Pecinka A, Schmidt R, McBreen K, Schubert I (2006) Mechanisms of chromosome number reduction in *Arabidopsis thaliana* and related Brassicaceae species. Proc Natl Acad Sci U S A 103:5224–5229
Maere S, De Bodt S, Raes J, Casneuf T, Montagu M, Kuiper M, Van de Peer Y (2005) Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci U S A 102:5454–5459
Mandáková T, Joly S, Krzywinksi M, Mummenhoff K, Lysak MA (2010) Fast diploidization in close mesopolyploid relatives of *Arabidopsis*. Plant Cell 22:2277–2290

Matsuno M, Compagnon V, Schoch GA, Schmitt M, Debayle D, Bassard J-E, Pollet B, Hehn A, Heintz D, Ullmann P, Lapierre C, Bernier F, Ehlting J, Werck-Reichhart D (2009) Evolution of a novel phenolic pathway for pollen development. Science 325:1688–1692

Mayrose I, Zhan SH, Rothfels CJ, Magnuson-Ford K, Barker MS, Rieseberg LH, Otto SP (2011) Recently formed polyploid plants diversify at lower rates. Science 333:1257

Mena M, Ambrose BA, Meeley RB, Briggs SP, Yanofsky MF, Schmidt RJ (1996) Diversification of C function activity in maize flower development. Science 274:1537–1540

Miyake T, Takebayashi N, Wolf DE (2009) Possible diversifying selection in the imprinted gene, *MEDEA,* in *Arabidopsis*. Mol Biol Evol 26:843–857

Ohno S (1970) Evolution by gene duplication. Springer, Berlin/Heidelberg/New York

Ohta T (1987) Simulating evolution by gene duplication. Genetics 115:207–213

Papp I, Mette MF, Aufsatz W, Daxinger L, Schauer SE, Ray A, van der Winden J, Matzke M, Matzke AJM (2003) Evidence for nuclear processing of plant micro RNA and short interfering RNA precursors. Plant Physiol 132:1382–1390

Pfeil BE, Schlueter JA, Shoemaker RC, Doyle JJ (2005) Placing paleopolyploidy in relation to taxon divergence: a phylogenetic analysis in legumes using 39 gene families. Syst Biol 54:441–454

Rizzon C, Ponger L, Gaut BS (2006) Striking similarities in the genomic distribution of tandemly arrayed genes in *Arabidopsis* and rice. PLoS Comput Biol 2:e115

Salse J, Abrouk M, Bolot S, Guilhot N, Courcelle E, Faraut T, Waugh R, Close TJ, Messing J, Feuillet C (2009) Reconstruction of monocotyledonous proto-chromosomes reveals faster evolution in plants than in animals. Proc Natl Acad Sci U S A 106:14908–14913

Sankoff D, Zheng C, Zhu Q (2010) The collapse of gene complement following whole genome duplication. BMC Genomics 11:313

Schmutz J et al (2010) Genome sequence of the paleopolyploid soybean. Nature 463:178–183

Schranz ME, Edger PP, Pires JC, van Dam N, Wheat CW (2011) Comparative genomics in the Brassicales: ancient genome duplications, glucosinolate diversification and Pierinae herbivore radiation. In: Edwards D, Batley J, Parkin I, Kole C (eds) Oilseeds: genetics, genomics, and breeding of oilseed brassicas. Science, Boca Raton

Schuermann D, Molinier J, Fritsch O, Hohn B (2005) The dual nature of homologous recombination in plants. Trends Genet 21:172–181

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

Senchina DS, Alvarez I, Cronn RC, Liu B, Rong J, Noyes RD, Paterson AH, Wing RA, Wilkind TA, Wendel J (2003) Rate variation among nuclear genes and the age of polyploidy in *Gossypium*. Mol Biol Evol 20:633–643

Shi T, Huang H, Barker MS (2010) Ancient genome duplications during the evolution of kiwifruit (*Actinidia*) and related Ericales. Ann Bot 106:497–504

Soltis DE, Soltis PS, Pires JC, Kovarik A, Tate JA, Mavrodiev E (2004) Recent and recurrent polyploidy in *Tragopogon* (Asteraceae): cytogenetic, genomic and genetic comparisons. Biol J Linn Soc 82:485–501

Spillane C, Schmid KJ, Laoueillé-Duprat S, Pien S, Escobar-Restrepo J-M, Baroux C, Gagliardini V, Page DR, Wolfe KH, Gossniklaus U (2007) Positive Darwinian selection at the imprinted *MEDEA* locus in plants. Nature 448:349–352

Stebbins GL (1950) Variation and evolution in plants. Columbia University Press, New York

Stebbins GL (1971) Chromosomal evolution in higher plants. Addison-Wesley, Reading

Stephens SG (1951) Possible significance of duplication in evolution. Adv Genet 4:247–265

Tang HB, Bowers JE, Wang XY, Ming R, Alam M, Paterson AH (2008) Synteny and collinearity in plant genomes. Science 320:486–488

Thomas BC, Pedersen B, Freeling M (2006) Following tetraploidy in an *Arabidopsis* ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome Res 16:934–946

Throude M, Bolot S, Bosio M, Pont C, Sarda X, Quraishi UM, Bourgis F, Lessard P, Rogowsky P, Ghesquiere A, Murigneux A, Charmet G, Perez P, Salse J (2008) Structure and expression analysis of rice paleo duplications. Nucleic Acids Res 37:1248–1259

Tian C, Xiong Y, Liu T, Sun S, Chan L, Chen M (2005) Evidence for an ancient whole-genome duplication event in rice and other cereals. Acta Genet Sin 32:519–527

Tuskan GA et al (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313:1596–1604

Veitia RA (2002) Exploring the etiology of haploinsuffinciency. Bioessays 24:175–184

Veitia RA (2003) Nonlinear effects in macromolecular assembly and dosage sensitivity. J Theor Biol 220:19–25

Veitia RA (2005) Paralogs in polyploids: one for all and all for one? Plant Cell 17:4–11

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

Wagner WH (1970) Biosystematics and evolutionary noise. Taxon 19:146–151

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

Wang X, Shi X, Hao B, Ge S, Luo J (2005) Duplication and DNA segmental loss in the rice genome: implications for diploidization. New Phytol 165:937–946

Wolfe KH (2001) Yesterday's polyploids and the mystery of diploidization. Nat Rev Genet 2:333–341

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

Woodhouse MR, Schnable JC, Pedersen BS, Lyons E, Lisch D, Subramaniam S, Freeling M (2010) Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homeologs. PLoS Biol 8(6):e1000409

Xiao H, Jiang N, Schaffner E, Stockinger EJ, van der Knaap E (2008) A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. Science 319:1527–1530

Xiong Z, Gaeta RT, Pires JC (2011) Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid Brassica napus. Proc Natl Acad Sci U S A 108:7908–7913

Yanofsky MF, Ma H, Bowman JL, Drews GN, Feldmann KA, Meyerowitz EM (1990) The protein encoded by the *Arabidopsis* homeotic gene *agamous* resembles transcription factors. Nature 346:35–39

Zou C, Lehti-Shiu MD, Thomashow M, Shiu S-H (2009) Evolution of stress-regulated gene expression in duplicate genes of *Arabidopsis thaliana*. PLoS Genet 5:e1000581