Polyploidy-associated genome modifications during land plant evolution

Yuannian Jiao and Andrew H. Paterson

Plant Genome Mapping Laboratory, University of Georgia, 111 Riverbend Road, Athens, GA 30606, USA

The occurrence of polyploidy in land plant evolution has led to an acceleration of genome modifications relative to other crown eukaryotes and is correlated with key innovations in plant evolution. Extensive genome resources provide for relating genomic changes to the origins of novel morphological and physiological features of plants. Ancestral gene contents for key nodes of the plant family tree are inferred. Pervasive polyploidy in angiosperms appears likely to be the major factor generating novel angiosperm genes and expanding some gene families. However, most gene families lose most duplicated copies in a quasi-neutral process, and a few families are actively selected for single-copy status. One of the great challenges of evolutionary genomics is to link genome modifications to speciation, diversification and the morphological and/or physiological innovations that collectively compose biodiversity. Rapid accumulation of genomic data and its ongoing investigation may greatly improve the resolution at which evolutionary approaches can contribute to the identification of specific genes responsible for particular innovations. The resulting, more ‘particulate’ understanding of plant evolution, may elevate to a new level fundamental knowledge of botanical diversity, including economically important traits in the crop plants that sustain humanity.

1. Introduction

Genome duplication is a punctuational event in the evolution of a lineage, with permanent consequences for all descendants—if the lineage survives. Most crown eukaryotes pass through different ploidy levels at different stages of development [1,2] and continuously produce aberrant unreduced gametes at low rates. However, the extreme rarity of genome duplications in the evolutionary history of extant lineages, usually surviving only once in many millions of years, shows that the vast majority quickly go extinct [3].

Classical views suggest that genome duplication is potentially advantageous as a source of genes with new functions [4,5]. Some polyploids appear to realize these and other benefits [6], with genome duplication thought to be central to the evolution of morphological complexity [7]. Polyploids have long been suggested to enjoy a variety of capabilities that transgress those of their diploid progenitors, such as adaptation to environmental extremes [8–11].

Angiosperms are an outstanding model in which to elucidate consequences of genome duplication in crown eukaryotes. It has long been suspected that many angiosperms were palaeopolyploids [10,12]. Indeed, recent analyses of genome sequences [13,14] show that all angiosperms are palaeopolyploids. Seminal findings from yeast [15–17] and Paramecium [18] are shedding valuable light on consequences of genome duplication in single-celled organisms. However, these consequences are expected to be very different in higher eukaryotes with small effective population sizes, such as angiosperms and mammals [19,20].

Herein, we describe some representative features of major land plant groups and associated genomic modifications and variations throughout the history of plant evolution. We circumscribe the ancestral gene content for key nodes of the evolutionary series, as well as adaptational genomic changes occurring along the branches leading to these nodes. We also review current approaches for genome structure comparison to unravel ancient polyploidy and discuss the consequences of these polyploidy events, particularly the biased pattern of gene gains and losses following genome duplication.
2. Genome resources in major plant groups

Land plants, also called embryophytes, descended from freshwater algae about 480 Ma [21–24]. Most of the major phylogenetic groups of land plants now have at least one high-quality draft or reference genome sequence. Mosses, with about 12,000 species classified in the Bryophyta, are non-vascular plants that lack xylem and absorb water and nutrients mainly through their leaves. To date, there is only one Bryophyta species (Physcomitrella patens) with a completed genome sequence [25]. Lycophytes, together with euphyllophytes (ferns and seed plants), are the two surviving lineages after the origin of vascular plants. Selaginella moellendorfii, in the lycophyte family Selaginellaceae (spike mosses), is the first vascular non-seed plant with its genome sequenced [26]. Seed plants (spermatophytes) include gymnosperms and angiosperms. Gymnosperms are a group of seed-producing plants; extant gymnosperms include conifers, cycads, Ginkgo and Gnetales. Many gymnosperms have exceptionally large genomes; for example, conifer genome sizes range from 18 to 35 gigabases [27], hindering whole genome sequencing. Picea abies is the first gymnosperm species with a full genome sequence, completed in 2013 [28]. Angiosperms are by far the largest group of land plants, with more than 300,000 living species, of which at least 33 have sequenced genomes (as in Phytotome v9.1), notably Arabidopsis [29], Populus [30], grape [31], Oryza [32], Sorghum [33], banana [34] and many more, with many additional genome sequences in progress. These genome sequences from species across land plant phylogeny provide critical resources to study genome evolution and associated innovations of plant form and function.

3. Ancestral genome modification at key evolutionary nodes

Clarification of the ancestral gene content and lineage-specific variations at key nodes in the land plant phylogeny will advance knowledge of how genomic modifications contributed to the evolution of novel features. Here, we present an exemplary study of reconstructing ancestral gene content. The protein-coding gene sets of 11 sequenced land plant species (five eudicots: Arabidopsis thaliana, Populus trichocarpa, Vitis vinifera, Solanum lycopersicum and Nelumbo nucifera; three monocots: Oryza sativa, Sorghum bicolor and Musa acuminata; one gymnosperm: Picea abies; one lycophyte: Selaginella moellendorfii; one moss: Physcomitrella patens) were used to identify putative gene family clusters by OrthoMCL [35]. The OrthoMCL approach builds tighter gene clusters with fewer false positives than the Tribe-MCL method [36]. However, the OrthoMCL clusters might not represent truly distinct gene families, but could be subfamilies for some more divergent gene families (gene family represents an OrthoMCL cluster as a general term hereafter). For example, the MADS-box gene family includes tens of OrthoMCL clusters. Therefore, we have to emphasize the potential over-estimation of ancestral gene content, as well as the number of gene family gains and losses. The number of genomes used in the gene family classification could also affect the gene clusters, especially since most of the currently available genomes are angiosperms. However, this analysis does illustrate a general trend of genome modifications following land plant evolution. We modelled the changes of gene content occurring along each branch using a Wagner parsimony framework model.

From this gene family classification, we estimated a minimum gene set of 7100 genes (figure 1, node 1) in the common ancestor of all land plants, which is very close to the estimation (6820) of Banks et al. [26]. Banks et al. [26] identified gene families using assignment of orthology by mutual-best-hit criteria and synteny between closely related organisms. The lycophyte Physcomitrella has more than twice the number of genes of aquatic Chlamydomonas (35,938 versus 15,143) [25,37], suggesting a general increase in gene family complexity following the transition to land. Genes associated with aquatic environments (e.g. flagellar components for gametic motility) and dynein-mediated transport have been lost in Physcomitrella [25]. By contrast, Physcomitrella gained members of gene families associated with signal transduction (e.g. through gibberellic acid, jasmonic acid, ethylene and brassinosteroids), transport capabilities and tolerance of abiotic stresses. Adaptation to land required plants to be able to survive much greater variations of temperature, light and water availability. An example of such adaptation may be the heat shock protein 70 (HSP70) gene family. All algal genomes sequenced to date have only one cytosolic HSP70 gene [38], while there are nine in P. patens [25]. Likewise, the light-harvesting complex proteins have significantly expanded in P. patens, perhaps contributing to robustness of the photosynthetic antenna to deal with high light intensities. Photo-protective early light-induced proteins also expanded greatly in P. patens, putatively associated with avoidance of photo-oxidative damage. Some of these expansions may be results of whole-genome duplications [25].

The estimated minimal gene number of the most recent common ancestor of vascular plants is 7790 (figure 1, node 2), with the acquisition of 690 new genes during the transition from non-vascular plants. In the lycophyte Selaginella, which, among plants with sequenced genomes, is sister to all other vascular plants, secondary metabolic genes (such as cytochrome P450, BAHD acyltransferases and terpene synthases) were expanded extensively [26]. It has been suggested that far fewer new genes were needed for the transition from a gametophyte to a sporophyte-dominated life cycle than for the transition from non-seed vascular plant to a seed plant and subsequently to a flowering plant [26].

A large number of gene family gains (1269) were inferred along the deep branch leading to seed plants, in which the estimated ancestral gene content was 8652. Genome size is exceptionally large for gymnosperms (18–35 gigabases), although polyploidy is thought to be rare in this group [39]. Recent efforts have indicated that the large genome size of gymnosperms might be associated with rapid expansion of retrotransposons and may be limited to conifers, Pinaceae [40–44], which are particularly well studied in view of their economic importance. A recent study suggested that elevated rates of genome size and diversification occurred within the past 100 Myr, especially in Pinus [45]. A draft sequence of the 20-gigabase genome of Picea abies has been recently completed [28]. Despite being more than 100 times larger than that of Arabidopsis, the Picea abies genome contains a similar number of well-supported annotated genes (28,354).

Comparison to Picea abies and/or other conifer genomes provides a valuable reference for studying genome modifications and the evolution of key traits for seed plants, for example the origin of flowering. The representative gymnosperm, Picea
family gains (Wagner parsimony framework. Numbers after each node indicate the estimated ancestral gene numbers, while the numbers above each branch show the number of gene families were identified using OrthoMCL, and the Count program was used to determine the minimum gene set for ancestral nodes of the phylogenetic tree using a

Figure 1. Global gene family loss and gain in plant genomes. Eleven sequenced plant genomes were selected for phylogenetic representation of major plant groups. Gene families were identified using OrthoMCL, and the Count program was used to determine the minimum gene set for ancestral nodes of the phylogenetic tree using a Wagner parsimony framework. Numbers after each node indicate the estimated ancestral gene numbers, while the numbers above each branch show the number of gene family gains (+) and losses (−). The extra bar under the internal branches means one or more rounds of ancient polyploidy suggested along that time period.

abies, lacks flowering locus T (FT)-like genes (that promote flowering in other taxa), instead containing a group of FT/TFL1-like genes which probably act as flowering repressors [28,46,47]. The FT-like genes are exclusively found in angiosperms, including early diverging and eudicot lineages, supporting the hypothesis that the evolution of flowering plants coincided with the evolution of a flower-promoting function for an FT/TFL1-like gene. In plants, MADS-box genes encode transcription factors, which are important regulators of plant development, particularly as regulators of floral organ identity [48]. A total of 278 MADS-box homologues were identified in P. abies, most of which are type II MADS-box genes (MIKC-type proteins, with structure of MADS (M) domain followed by an Intervening (I), a Keratin-like (K) and a C-terminal domain). The VASCULAR NAC DOMAIN (VND) gene family controls the formation of multi-cellular vessels, which has been considered to be one of the key innovations that contribute to the success of the flowering plants [49]. The VND gene family has been expanded with only two VND genes found in P. abies [28] and seven in Arabidopsis thaliana [50].

Another large wave of gene family gains (figure 1, node 4, 1492 new gene families) was inferred along the branch leading to angiosperms, suggesting that a diverse set of novel gene functions originated before the emergence of angiosperms. Angiosperms have seeds contained within a fruit, unlike gymnosperms that have naked seeds (no fruit). The flowers, fruits and other characters of angiosperms are likely to have contributed to their emergence as the most species-rich group of land plants [51–53]. However, because this exemplar analysis does not include genome sequences from basal angiosperms, it cannot distinguish gene families originating before or soon after the earliest diversification of angiosperms. Amborella is thought to be the single living sister species to other extant flowering plants, and the Amborella trichopoda genome, which is finished but not yet available for large-scale analysis, will provide a valuable reference for reconstructing ancestral angiosperm gene content (http://www.amborella.org/) [54]. The estimated minimum gene sets for eudicots and monocots are 10 994 and 10 547 (figure 1, nodes 5 and 9), respectively. A total of 1068 novel gene families were gained in eudicots as a whole (on the branch to node 5), and 662 of these were gained before the diversification of monocots (on the branch to node 9). However, the largest gain of gene families in the internal branches of the studied 11-genome phylogeny was inferred along the branch leading to grasses (figure 1, node 10), with about 4126 gene families newly gained before the divergence of Oryza and Sorghum.

4. Genome structure comparison and synteny detection

Comparisons between and within genomes permit one to detect homologous regions based on conserved gene content and gene order (synteny blocks), in some cases also revealing that large-scale, or whole-genome, duplications (WGDs) occurred during evolutionary history. In vertebrates, synteny blocks are still detectable even after hundreds of millions of years of divergence [55,56]. However, angiosperm genomes are more dynamic with much more rapid structural evolution, which is due partly if not largely to extensive chromosomal rearrangement and gene loss following WGDs in the ancestral lineages [57–59]. For example, the two main clades of angiosperms, monocots and eudicots, were predicted to have diverged about 140–150 Ma [60]. Syntenic orthologue pairs between monocot and eudicot species can only account for 1–7% of the
total orthologue pairs, versus 30–50% in intra-species comparison in eudicots or monocots (see table 1 in [61]). Inter-genome synteny blocks can be substantially affected by the number of genome duplications after species divergence, as well as the timing of species divergence (figure 2).

Many algorithms and software have been developed to find syntenic blocks within or among genomes. In general, homology matrices are built using all-against-all BLAST searches, and then synteny is detected through clustering neighbouring matches within a matrix, such as the methods implemented in

\[ \text{no. WGDs after divergence increasing} \]

\[ \text{species divergence time increasing} \]
ADHoRe [62] and DiagHunter [63]. Another method to detect synteny is to employ dynamic programming to find the highest scoring ‘chains’ of syntenic genes, performing empirical or statistical tests to determine whether the observed synteny could have arisen by chance as implemented in DAGchainer [64], ColinearScan [65], MCscan [58] and MCscanX [61]. In addition, several web-based systems have been developed for whole-genome analysis. For example, CoGe is a platform for multiple whole-genome comparisons to find and align syntenic regions and visualize the output in an intuitive and interactive manner [66–68]. At the time of writing, it has implemented 15 140 organisms with 19 624 genomes, including Bacteria, Archaea, eukaryotes, organelles, viruses and sub-genomes such as plasmids. The Plant Genome Duplication Database (http://chibba.pgm.n.uga.edu/duplication) is a public web service to identify synteny information among plant genomes and is mainly focused on unravelling genome duplication events during the history of angiosperm evolution [69].

While syntenic blocks are the most definitive evidence supporting the most recent WGD in a genome, most angiosperm genomes have experienced multiple WGDs. In order to find the signatures of more ancient events, two approaches have been successfully used to track the ancient WGDs, referred to as ‘bottom-up’ [57] and ‘top-down’ approaches [58]. The ‘bottom-up’ approach is to ‘merge’ regions duplicated in the most recent event affecting a genome, using genes that remain duplicated and in syntenic locations to align corresponding regions and infer the gene content and order of an ancestral chromosome that would be sufficient to explain their gene content and order. Deduced ancestral chromosomes are then compared to one another in the same manner, to recursively reconstruct older genome duplications in a stepwise fashion [57,70]. The main challenge is how to order the non-anchor genes located in the syntenic blocks. To date, only the second and the third iteration have been achieved using the bottom-up approach [57,70]. In the ‘top-down’ approach, pair-wise syntenic blocks are first identified between the target genome and an outgroup genome. Then, all homologous segments between the two genomes are clustered together to infer the number of WGDs based on the redundancy level, i.e. the number of homologous segments each corresponding to a single region of the genome [13]. For example, in the tiny Utricularia gibba genome (genome size of 82 megabases), a top-down approach could identify at least three rounds of WGD occurring after the divergence of U. gibba from a common ancestor shared with tomato and grape [71].

5. Palaeopolyploidy and gene family numbers

Gene doubling may confer a number of advantages to a polyploid [6], via mechanisms such as increased gene dosage, ‘intergenomic heterosis’ conferred by multiple alleles in a polyploid nucleus, or the evolution of novel gene functions (neofunctionalization, [4,5]). Each of these advantages requires that each of the two duplicated copies of a gene survive in the doubled genome, thus increasing gene family numbers. However, it is apparent from a host of genome sequences that by far the most common fate of a pair of duplicated genes is silencing (either epigenetically or by mutation) and eventual elimination of one member of a pair, called ‘non-functionalization’. Classical ideas about neofunctionalization as a major advantage of polyploidy [4,5] have more recently been tempered with the suggestion that a more widespread fate of those duplicated genes that survive may be the evolution of subdivisions of ancestral functions (subfunctionalization, [72]) that render them interdependent. Subfunctionalization may sometimes lead to neofunctionalization [73].

In angiosperms (which are certainly representative of crown eukaryotes, and probably of all organisms in this regard), we find three ‘fates’ of individual gene pairs following duplication:

1. Most gene functional groups show post-duplication gene preservation/loss rates that are indistinguishable from the genome-wide average. Such ‘neutral’ loss of duplicated genes presumably involves inactivating mutations opposed by very weak selection [74], closely resembling the ‘non-functionalization’ described by others (e.g. [75]) as the fate of the vast majority of duplicated genes.

2. Genes in some specific functional categories tend to be retained after duplication. Several gene functional groups are preferentially preserved in duplicate [58,76–78]. Coding regions of genes preserved in duplicate tend to be functionally complex [76], under purifying selection [75,76] and evolve in concert [79,80]. Tendencies to retain duplicate genes involved in signal transduction and transcription and to lose DNA repair genes are widely observed [81,82]. However, Pfam domain-based groupings reveal heterogeneity in the broader and more widely used gene ontology (GO) categories used in prior studies [81,83], for example, showing one abundant protein–protein interaction domain (LRR, leucine-rich repeat) to be usually preserved in duplicate while two less-abundant domains (SET, TPR, tetratricopeptide repeat) are usually restored to singleton states [78].

3. Other specific genes and gene functional groups show more extensive loss of duplicate copies than the genome-wide average, and this loss is often convergent following independent duplications separated by hundreds of millions of years (i.e. in angiosperms, yeast and the fish Tetrasporon [78]). Some gene functional groups are preserved in duplicate significantly less frequently than the genome-wide average. This observation alone might be viewed as noise—among thousands of functional groups, some must incur more gene loss than others due to random factors. However, the gene functional groups that have incurred the greatest loss of duplicated copies are closely correlated following independent duplications in Arabidopsis and rice, at statistical probabilities that essentially rule out false positives [78]. Multi-alignments show some individual genes to have been repeatedly restored to single-copy status following many different genome duplications in independent angiosperm lineages [13,58]. The two gene sets do not overlap. Repeated restoration of certain genes to singleton status at a greater-than-random frequency suggests that an underlying set of principles of molecular evolution may contribute to the fates of gene and genome duplications [78].

In the following pages, we elaborate in more detail on these respective outcomes.

6. Palaeopolyploidy and gene family gains

Gene duplication has long been thought to be a primary source of material for evolution [4]. As discussed above, the duplicates retained in genomes usually have divergent functions. A now-classical concept is that a duplicate gene is relatively free of selective constraints and can evolve new
functions (neofunctionalization), as long as the ancestral copy maintains its original functions [4,84]. Thus, gene duplication has been viewed as one of the main molecular mechanisms in the creation of new genes [85]. Single gene duplicates often lose part or all of the regulatory features of their progenitor genes, thus being prone to evolving new expression patterns and perhaps new functions [86]. However, single-gene duplicates have relatively short half-lives [84], so evolution of a new function that confers a selective advantage needs to happen relatively quickly or the gene may be lost. Under WGD, involving the doubling of the entire genome, duplicated genes retain their ancestral regulatory features and expression patterns at least initially. However, they generally survive longer than single gene duplicates and thus have more time to evolve novel function that justifies their retention, in principle offering a rich potential source of new genes, gene families and sub-families in addition to increasing the number of members of existing gene families.

The rapidly expanding set of whole genome sequences available for plant species has provided useful resources for unravelling palaeopolyploidy events in evolutionary history [13,33,57,83,67], as well as for reconstructing ancestral gene content. We investigated gene family gains along deep branches of the land plant phylogeny, and found that the largest number of gene family gains is on the branches where polyploidy events occurred (figure 1). As mentioned above, these gains could mean true new gene family origination, or just expansion of extensively diverged subfamilies. Ancient polyploidy events in the common ancestors of all extant seed plants, and all extant angiosperms, respectively, were identified using a genome-scale phylogenomic approach [87]. These two ancient WGD events appear on the branches leading to nodes 3 and 4 in figure 1, respectively, on which large numbers of gene family gains occurred (1269 and 1492). The origin of the seed and flower may be partly explained by these novel genes and expanded gene families that appeared before the diversification of seed plants and angiosperms. For example, 35 floral genes in Arabidopsis were clustered in orthogroups that originated before the angiosperm radiation (electronic supplementary material, table S1). Many floral genes originated before seed plants and have been further expanded in number by ancient polyploidy events, such as MADS-box [28], PHYTOCHROME [88] and HD-ZIP III gene families [89]. In total, we found 21 genes associated with flowers in the orthogroups newly gained before the divergence of angiosperms and gymnosperms (electronic supplementary material, table S2). These results suggest that gene recruitment is an ongoing process, while gene origination or expansion could be a short/long time before the recruitment. Indeed, gene recruitment to a function surely requires not only available genes but also an external impetus, such as climate change. For example, duplicate copies of virtually all genes that were eventually recruited into the evolution of C4 photosynthesis in Poaceae grasses were available following a pan-Poaceae genome duplication at about 70 Ma [90]. However, many of these copies were lost, and subsequent single-gene duplicates of the same genes were recruited into C4 photosynthesis about 40 Myr later [91]. The largest gene family gain is 4126, on the internal branch that gave rise to Oryza and Sorghum. At least two (ρ and σ), and perhaps three, WGD events were identified on this branch leading to node 10, which is specific to the Poaceae (grass) clade, not shared with the Musa (grass) lineage [34,70]. By examining syntenic evidence, the most recent WGD, p, was identified as a shared event of Oryza and Sorghum [90]. A bottom-up reconstruction approach was used to build pre-p ancestral blocks. A second iteration of synteny analysis was able to find a more ancient polyploidy event before p [70]. Genome comparisons revealed that generally three regions of the eudicot Vitis (grape) genome matched up to eight homologous regions in Oryza, suggesting that three WGD events occurred in the Oryza lineage after the split with eudicots [70]. Another ancient polyploidy event on the internal branches, named gamma, is most likely shared with all core eudicots [31,92–94]. There are 623 novel gene family gains along the branch leading to core eudicots (node 6). However, more gene family gains (1068) appear on the branch leading to all eudicots (node 5), where no genome duplication has been identified. This is somewhat in conflict with the other large gains, which usually accompanied one or more WGD events. The inconsistency could be due to the timing and nature of the gamma event, which may actually be two events that occurred in close proximity, very early in the evolutionary history of eudicots [93].

Changes in gene family content along terminal branches of the land plant phylogeny could be partly due to further independent polyploidy events in the evolutionary history of many lineages. At least one more polyploidy event was identified along the branches leading to all eudicot lineages depicted except Vitis [30,58,93,95]. Three independent polyploidy events were proposed on the terminal branch to Musa acuminata [34]. Numbers of gains inferred along the terminal branches could also be somewhat inflated by genome annotation errors.

Polyploidy events appear likely to be the major factor generating novel genes, as well as expanding gene families. For angiosperm branches on which no ancient polyploidy event was identified (branches leading to nodes 2, 7, 8 and 9), relatively small numbers of novel gene families were gained (690, 213, 434 and 662, respectively). Small-scale duplications, such as transposition and tandem duplications, might contribute these gene family gains. Indeed, while polyploidy may confer the largest number of new genes, small-scale duplications that separate genes from ancestral regulatory features and place them in new genomic environments, are more prone to the evolution of novel expression patterns and may be more likely than polyploidy-derived genes to give rise to novel functions [96].

It has been widely acknowledged that gene retention after polyploidy events is biased [97–100]. Genes in functional categories of transcription factors, protein kinases and transferases, tend to retain duplicated copies following WGD [81,97]. Several models have been proposed to explain the biased retention pattern following duplication. Observations in Saccharomyces cerevisiae and Caenorhabditis elegans [101] suggest that genes for which duplicates are preserved tend to be slowly evolving, whereas genes restored to singletons tend to have a higher mutation rate. Another idea relates gene retention probability to gene expression level. In yeast, it has been observed that selection for increased levels of gene expression was a significant factor determining which genes were retained and which were returned to a single-copy state [102]. In addition, the ‘gene balance’ hypothesis [97] predicts a relationship between the number of interactions, or ‘connectivity’ of a gene (i.e. with other genes), and its sensitivity to altered gene dosage. ‘Connected’
genes are more likely to be retained after WGD than after tandem duplications, with losses of such genes resulting in a sort of haploinsufficiency relative to the remainder of the genome. It has also been argued that balanced gene drive should tend to drive up morphological complexity [7], which is potentiated by duplicated functional models (gene

Figure 3. Different gene loss patterns affect the reconstructed phylogeny using single-copy genes. If the single-copy gene was duplicated in the common ancestor of species A, B and C, different gene silencing timing and patterns will impact the final reconstructed evolutionary relationships. The most hypothesized case is that the duplicates in the ancestor are quickly restored to single-copy status. However, several single-copy genes in soya bean have retained the duplicated copy for about 13 Myr. Many speciation events could occur during such a long time. Here, we proposed four different gene loss patterns, and all could restore single-copy gene status. (a) All three genes in one big clade went through gene deletion, and the remaining three genes can be used to reconstruct the phylogeny correctly. However, if gene loss patterns were as in (b) and (c), the constructed phylogeny will not truly reflect the relationships among species. (d) Although gene c2 is not the orthologue of a1 and b1, the final tree is correct by chance. (Online version in colour.)
networks). An important need for further investigation of many of these ideas is a much greater knowledge of the various forms of interactions among genes, including physical, regulatory and feedback-mediated.

7. Palaeopolyploidy and gene losses

Polyploidy has made a significant contribution to gene family expansion and novel gene content as discussed above; however, non-random patterns of loss of duplicated genes are less studied. As noted above, the vast majority of gene duplicates tend to be silenced in a relatively short period of time, and this process is not random [81,103,104]. Speciﬁc genes and gene functional groups show more extensive loss of duplicate copies than the genome-wide average, and this loss is often convergent following independent duplications separated by hundreds of millions of years [100,105,106]. Investigations of single-copy genes in sequenced flowering plants has statistically ruled out the possibility of random gene loss (e.g. [58,106]). Single-copy genes are overrepresented in some essential functional categories, such as DNA repair, recombination, enzyme activity and organelle-related functions, and underrepresented in GO categories in which WGD survivors are usually signiﬁcantly enriched (e.g. transcription factor activity, kinase activity, transport, signal transduction) [105,106]. Furthermore, single-copy genes show more sequence conservation, higher gene expression level, and expression in more tissues than multiple-copy genes [106]. Conserved nuclear single-copy genes have been used as markers for reconstructing angiosperm phylogeny, even eukaryotic relationships, and for improving resolution of the evolutionary history of organisms [107,108]. However, it might be problematic to use some specific single-copy gene families as phylogenetic markers if speciation precedes their restoration to single-copy status. For example, two recent studies [106,107] each show that genes that are single-copy in other organisms are still duplicated in soya bean after its most recent polyploidization event at approximately 13 Ma, and tetraploid cotton appears likewise to experience relatively slow gene loss [109]. If a speciation event occurred before duplicated genes experienced reciprocal gene silencing, a reconstructed phylogeny may not reﬂect the true evolutionary relationship (ﬁgure 3).

It has been suggested that reciprocal gene loss after polyploidy could contribute directly to speciation and reproductive isolation [17,90,110–112]. By investigating gene losses following a WGD in a common ancestor of three yeast species (S. cerevisiae, Saccharomyces castellii and Candida glabrata), it has been shown that 20% of the loci experienced differential gene loss patterns [17]. The speciations were shortly after the WGD event, during a period of precipitous gene loss. Therefore, it is hypothesized that reciprocal gene loss at many ancestrally duplicated genes could be the main factor leading to speciation [17]. However, two recent genome-wide studies of syntenic gene losses in Poaceae found very little evidence supporting reciprocal loss of homologous genes among the grass species [33,113]. While evidence from additional plant groups needs to be explored, substantially different effective population sizes in microbes (very large) and crown eukaryotes (very small) may contribute to differences in the fates of duplicated genes [19,20] and associated evolutionary patterns.

8. Conclusion and future studies

One of the great challenges of evolutionary genomics is linking genome modiﬁcations that are evident in the burgeoning sets of angiosperm (and other) genome sequences now available to speciation, diversiﬁcation and the morphological and/or physiological innovations that collectively constitute biodiversity. Polyploidy events are prevalent throughout angiosperm evolution. While the timing of many ancient WGD events can be circumscribed to speciﬁc branches of angiosperm phylogeny, much better resolution is needed to link genomic changes to their biological consequences. Soon, genome sequences will be available for virtually all branches of land plant phylogeny, which will help to improve resolution. Enriched knowledge of botanical diversity may permit us to circumscribe not merely thousands or even hundreds of important functional changes to a branch, but much smaller numbers, approaching a more ‘particulate’ model for evolutionary history in much the same manner that Mendelian genetics transitioned science away from ‘blending’ models of inheritance. Dosage balance could explain some biases in gene retention after different modes of duplication. Detailed analysis of evolutionary history of all members of functional protein complexes might provide stronger support for this hypothesis. More effort is needed in reconstructing, decoding and analysing the inferred ancestral genomes prior to diversiﬁcation of major angiosperm groups. With reconstructed ancestral genomes, we could thoroughly assess genome modiﬁcations such as gene gain and loss through evolution, and associate such changes with diversiﬁcation and novel innovation and function. In addition, genome resequencing data could provide more insights into genome variation following evolution and adaptation at the population level. Genome modiﬁcations, including duplication, fractionation, rearrangements and changes in expression, have the potential to facilitate the origin of new functional variation in organisms, including economically important traits in the crop plants that sustain humanity.

References


60. Chaw SM, Chang CC, Chen HL, Li WH. 2004 Dating genomic sequencing for classical genetics in E. coli (doi:10.1007/s12042-008-9017-y)


