Review

The flowering world: a tale of duplications

Yves Van de Peer[1,2], Jeffrey A. Fawcett[1,2], Sebastian Proost[1,2], Lieven Sterck[1,2] and Klaas Vandepoele[1,2]

[1] Department of Plant Systems Biology, Flanders Institute for Biotechnology (VIB), 9052 Gent, Belgium
[2] Department of Molecular Genetics, Ghent University, 9052 Gent, Belgium

Flowering plants contain many genes, most of which were created during the past 200 or so million years through small- and large-scale duplications. Paleo-polyploidy events, in particular, have been the subject of much recent research. There is a growing consensus that one or more genome doubling or merging events occurred early during the evolution of the flowering plants, and that many lineages have since undergone additional, independent and more recent duplication events. Here, we review the difficulties in determining the number of genome duplications and discuss how the completion of some additional genome sequences of species occupying key phylogenetic positions has led to a better understanding of the timing of certain duplication events. This is important if we want to demonstrate the significance of genome duplications for the evolution and radiation of (different groups of) flowering plants.

Flower power

It is hard to imagine a world without flowering plants. They represent one of the greatest radiations in the history of life on Earth, with >350 000 known species. However, in spite of much research, the origin of this large clade of organisms remains unclear. Although seed plants appeared during the Late Devonian, ~370 million years ago (mya), the flowering plants, or angiosperms, appeared later in the fossil record, during the Early Cretaceous, ~130–135 mya [1]. The major lineages of angiosperms, including those of monocots and eudicots (see Glossary), might have diverged within a few million years of each other, perhaps ~140–150 mya [2]. By 90 mya, most of the extant major lineages were established, and angiosperms had become the most dominant lineage on Earth [3].

It is generally well acknowledged that the origin and radiation of the flowering plants is not old (relatively speaking, for instance, compared with the origin of the green lineage, which is more than a billion years ago) and occurred relatively abruptly [4]. The question remains over what triggered and explains the important changes in the evolution of plants during the Late Jurassic – Early Cretaceous, and made the flowering plants one of the most successful groups of organisms on the planet [5,6]. 'An abominable mystery' was how Charles Darwin referred to this rise and fast diversification of the angiosperms, and the puzzle remains as controversial as ever. Analysis of the ever-increasing amount of genomic data suggests that part of the answer for the origin and evolutionary success of the flowering plants might lie in gene duplication, but perhaps even more so, in whole-genome duplications (WGDs). Here, we discuss how remnants of such WGDs can be unveiled and dated.

Whole-genome duplication early in the evolution of flowering plants

Analyses based on whole genome sequences revealed that angiosperm species show traces of genome doublings early in their evolution (see Box 1). For instance, it was initially suggested that one of the WGDs detected in *Arabidopsis thaliana* (commonly referred to as the *beta* duplication) occurred before the radiation of most eudicots, and that the oldest WGD detected in *Arabidopsis* (the *gamma* duplication) is probably shared with *Oryza sativa* (rice) [7,8]. Analyses of gene families, such as various MADS-box

Glossary

Co-sexual flowers: have both male and female reproductive structures, including stamens, carpels and an ovary.
Eudicots: or tricolpates, literally, 'true dicotyledons', contains most plants that have been considered dicotyledons. Dicotyledons, or dicots, refer to a non-monophyletic group of flowering plants whose seed typically contains two embryonic leaves or cotyledons.
Heterosis (or hybrid vigor): the crossing of two in-bred lines results in progeny that are more vigorous than either parental line.
Monocots: a group of flowering plants whose seed typically contains one embryonic leaf or cotyledon.
Penalized likelihood methods: weigh the likelihood model of branch lengths in trees by a penalty function applying data-independent constraints to the parameter values. When used for dating purposes, these methods attempt to account for rate variation by penalizing rates that change too quickly from a branch to a neighboring branch. It is considered as an intermediate between molecular clock and completely unconstrained rate variation.
Polyploidy: a polyploid organism has more than two sets of chromosomes. Whereas autopolyploidy refers to a simple doubling of a single genome, an allopolyploidy event involves the merging of two differentiated, although similar, genomes. In allopolyploids, these duplicated (homoeologous) copies of a particular chromosome (segment) are derived from different donor taxa at the time of polyploid formation. For paleopolyploidy, see whole-genome duplication.
Synteny: originally defined as gene loci located on the same chromosome. This term has been used in comparative analyses to refer to chromosomal segments or to gene loci in different organisms located on a chromosomal region originating from a common ancestor. Genomic regions with collinear genes (i.e. showing conserved gene content and order) are often referred to as syntenous or syntenic.
Transitive homology: when homology between two genomic segments can only be inferred through a third gene or segment.
Whole-genome duplication (WGD): the simultaneous acquisition of extra copies of all of the nuclear chromosomes of an organism. Recent and ancient WGD events are also called polyploidy and paleopolyploidy, respectively.

Corresponding author: Van de Peer, Y. (Yves.vandepeer@psb.vib-ugent.be).

1360-1385/$ – see front matter doi:10.1016/j.tplants.2009.09.001 Available online 7 October 2009

Box 1. The ancestral angiosperm genome

Plant genomes contain many genes. More than 27 200 protein-encoding genes have been predicted for *Arabidopsis* [47], ~45 000 for poplar [30], >34 000 for *Lotus* [48], ~40 000 for *Medicago truncatula* [41], ~25 500 genes for papaya [13] and ~30 000 genes for grapevine [12,18]. For rice, the estimated number of ~42 500 non-transposon related genes was recently reduced to ~31 500 [49,50]. These numbers should all be interpreted with caution because many are based on first rounds of genome annotations and are likely to change with future annotation efforts. Nevertheless, numbers are generally high and differ substantially among species.

These differences are, apart from the number of large-scale duplication events, likely to be caused by differences in generation time. For instance, although *Arabidopsis* has, as far as we can tell, undergone the most genome doublings of all plants described here (hexapolyploidy followed by two additional WGDs), it has the fewest protein-encoding genes, taking the preliminary numbers of genes in other genomes at face value. This means that *Arabidopsis* must have lost a larger set of genes, and there are some indications that this might be the case. For example, there is a considerable number of gene families that are present in poplar, *Medicago* and rice, but not in *Arabidopsis* [49]. Furthermore, it has been shown that *Arabidopsis* has a faster substitution rate than do many other species [17,29], which also means that genes following duplication might turn into pseudogenes faster and eventually are lost. By contrast, the large number of genes in poplar could be explained by its slower rate of evolution. Whereas *Arabidopsis* is an annual weed, a single poplar (tree) genotype can persist as a clone for hundreds or even thousands of years, and recurrent contributions of ancient gametes from old individuals could account for the markedly reduced rate of sequence evolution, and thus, also gene loss [30]. Consequently, this would also imply that many poplar genes are still on the track to pseudogenization.

Based on a mathematical model that simulates the birth and death of genes through small- and large-scale gene duplication events, it has been estimated that the different polyploidy events in *Arabidopsis* have been directly responsible for >60% of the total number of duplicates that have been retained during the past ~150 million years [10]. From this number, it was inferred that the ancestral angiosperm genome contained no more than ~14 000 genes. Although this is based on the analysis of a single genome, similar values have been obtained through the comparison of different genomes. For instance, comparing the *Arabidopsis* and poplar gene sets suggests an ancestral gene count of 12 000 [30], whereas clustering of homologous genes from *Arabidopsis*, rice and 32 other plant species delineated ~12 400 ancestral genes [51]. Recently, counting the number of genes that show cross-species synteny between the genomes of *Arabidopsis*, grapevine, papaya and poplar, suggested 10 000–13 000 ancestral angiosperm genes [17].

It is clear that the number of genes in angiosperms has expanded during the past 150 million years or so. It has also been shown that, following WGDs, genes associated with transcription, signaling and development, in particular, have been retained [17,52–55]. Furthermore, the biased retention of regulatory genes has also been demonstrated in fungi and animals [56,57]. By contrast, few regulatory and developmental genes appear to have survived small-scale duplication events, in which individual genes have been duplicated. This is in agreement with the 'gene balance' hypothesis, which states that retention of genes that might have strong dosage effects, such as transcription factors, will be selected against if they are copied without their partners in the regulatory or protein interaction network [6]. However, if the genes encoding products that cooperate in the same complex pathway or network are duplicated simultaneously, which is the case in WGDs, gene dosage effects might be avoided by retaining all genes in that particular complex or network [10,58]. In addition, the products of regulatory genes are often active as protein complexes and need to be present in stoichiometric quantities for their correct functioning, which again is congruent with their high retention rate following whole-genome instead of small-scale duplication events [59,60].

transcription factor genes, also revealed the duplication of many of these genes early in the evolution of angiosperms through WGD, and it was suggested that the duplication and diversification of these genes was crucial to the radiation and success of angiosperms [4,9–11]. The recent completion of the genomic sequencing of *Vitis vinifera* (grapevine) and *Carica papaya* (papaya) revealed new evidence regarding the number and timing of WGDs that occurred early in the history of angiosperms, which contradicted previous interpretations [12,13]. *Vitis* is thought to be an early-diverging rosid, sister to both *Arabidopsis* and *Populus* (poplar), a hypothesis that has been supported by several recent large-scale phylogenetic studies [2,14–16]. Although a common ancestry of grapevine and poplar, and a sister-group relationship of *Arabidopsis* to grapevine and poplar were assumed based on overall protein sequence similarity [12], this close relationship between poplar and grapevine probably reflects the (much) faster evolutionary rate of *Arabidopsis* compared with poplar and grapevine, rather than their true evolutionary relationship [17,18]. Regions in the grapevine genome typically show homology with two other regions elsewhere in the genome, representing a triplicate structure (Figure 1). It was therefore concluded that three ancestral genomes had contributed to the grapevine lineage [12]. The recently released papaya genome shows a similar triplicate structure [13], although papaya is not closely related to grapevine. Instead, it belongs to the order Brassicales, and was estimated to have diverged from *Arabidopsis* only ~70 mya [13,14]. The most parsimonious explanation would be that the hexaploid origin (leading to a triplicate genome structure) is ancestral and shared between grapevine and papaya. Additional duplications in *Arabidopsis* appear not to be shared with papaya, meaning that the *Arabidopsis* lineage underwent two genome duplications (*alpha* and *beta*) after its divergence from the papaya lineage (see below, and Figure 2).

The exact timing and nature of the events giving rise to the ancestral triplicate genome structure is still debated [5]. For instance, although there is some evidence for an older duplication in the rice genome [19–21] (Figure 1), conclusive evidence that the hexaploidy in eudicots has been shared with monocots is lacking [12,17]. A genome duplication has also been proposed, based on extensive collections of Expressed Sequence Tag (EST) data, early in the evolution of magnoliids in the common ancestor of tulip poplar (*Liriodendron tulipifera*) and avocado (*Persea americana*) [22], at least 100 mya [23]. However, there is no genomic assembly for any magnoliid species yet, and it has not been investigated whether this large-scale gene duplication event corresponds to the hexaploidy event or represents an independent WGD in the magnoliid lineage. Thus, although some monocot and magnoliid species do show traces of duplications early in their evolution, it is yet to be established whether the hexaploidy is shared with monocots and magnoliids, and if not, which and how many lineages share these old duplications proposed in these groups. What is clear, however, based on a recent analysis

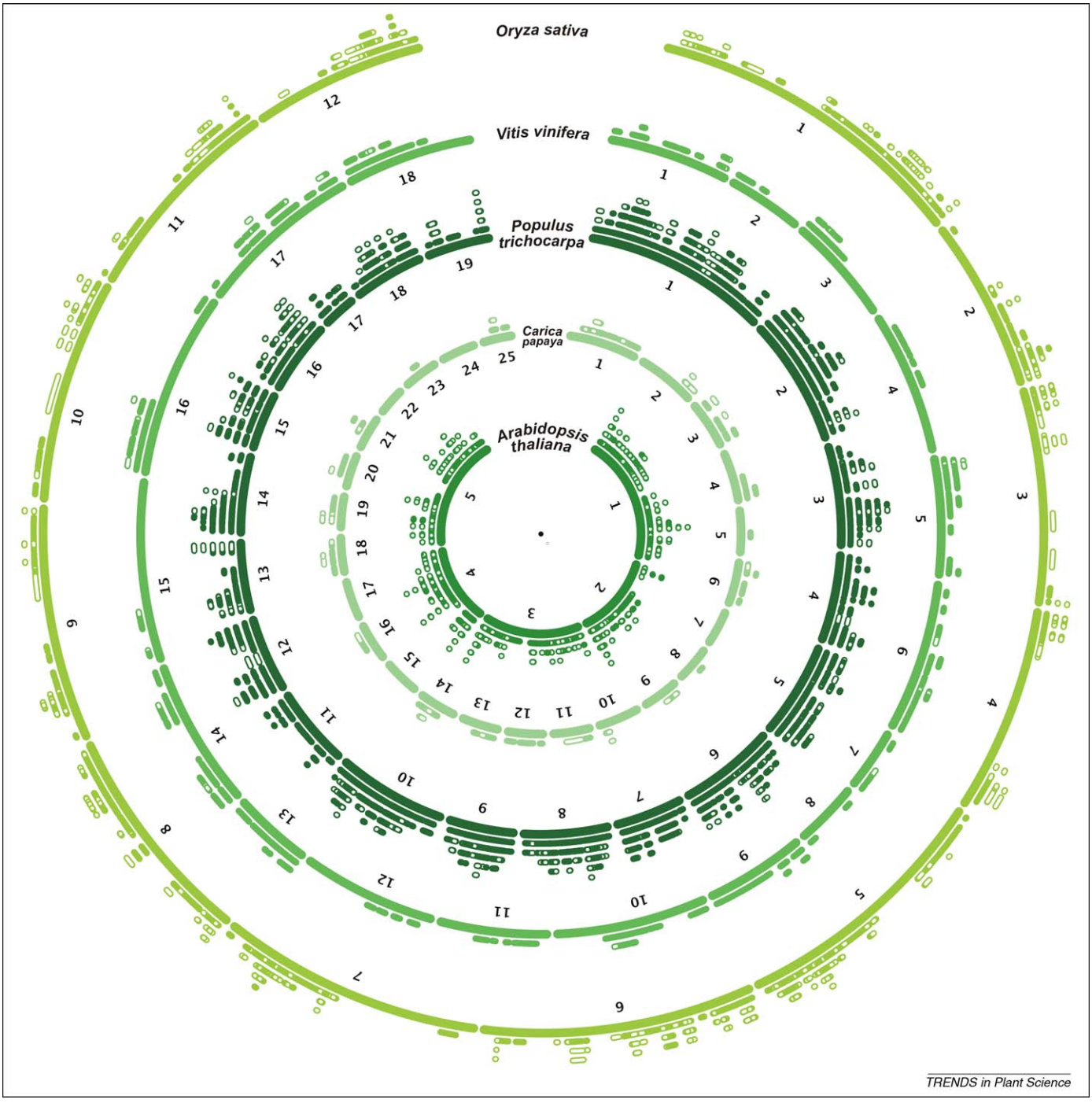

Figure 1. Circle plot showing intra-genomic homology for *Arabidopsis thaliana*, papaya (*Carica papaya*), poplar (*Populus trichocarpa*), grapevine *(Vitis vinifera*) and rice (*Oryza sativa*). Chromosomes are ordered clockwise. In the case of papaya, only the 25 largest scaffolds, in terms of number of genes, are depicted. Stacks (containing at least 30 genes) denote the segments that are homologous (i.e. duplicated) to that particular chromosomal location. Parts from a stack that are shown as being hollow represent duplicated segments that could only be uncovered through the use of transitive homology or genomic profiles [35,37]. Filled boxes represent segments for which the homology is obvious. The triplicate structure for both grapevine and papaya is confirmed [12,13], as is the sextuplicate structure for poplar. The multiplication level for *Arabidopsis* is less clear. For rice, there is also evidence for multiple duplications (see main text for details).

considering collinearity between tomato (*Solanum lycopersicum*) and the triplicate regions in grapevine, is that the hexaploidy occurred before the split between asterids and rosids and therefore pre-dates the divergence of most eudicot lineages [17]. Traces of ancient polyploidy were also observed in the EST data of several Asteraceae species, which might correspond to the hexaploidy event [24]. To investigate whether the triplicated structure in grapevine is the result of a step-wise genome merging process, phylogenetic tree analysis of homologous gene triplets located on the three grapevine subgenomes was performed, but no dominant tree topology was observed, indicating that the subgenomes are closely related [17]. However, by comparing the pattern of gene loss of paralogous segments, it was observed that two of the three subgenomes were more fractionated [25]. Thus, it was suggested that a first duplication event generated a tetraploid, which crossed with a diploid to generate a triploid,

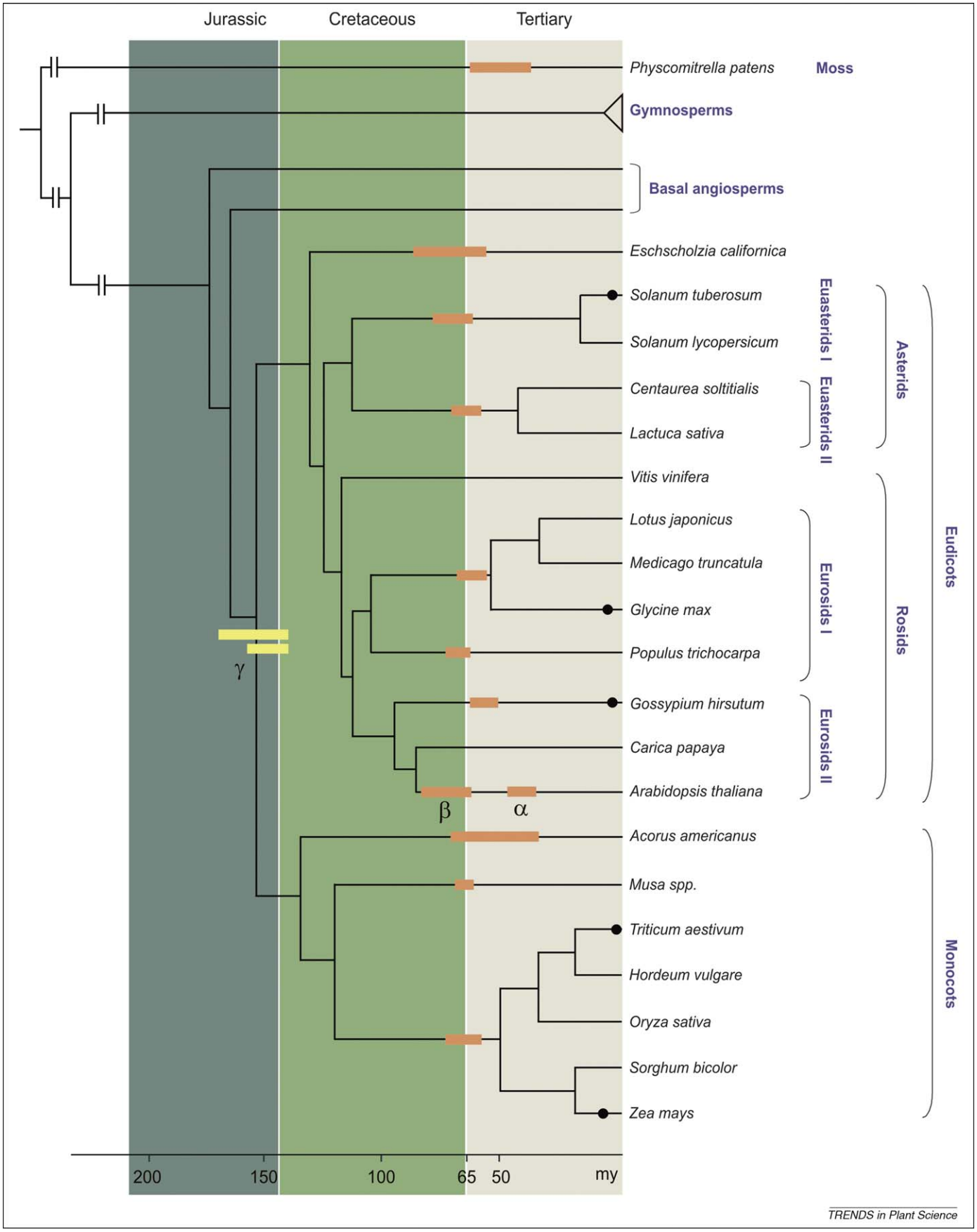

Figure 2. Phylogenetic tree of flowering plants (eudicots and monocots). WGDs, inferred from recent studies [28–30], are indicated by horizontal bars. Yellow bars denote the hexaploidy event. More recent WGDs appear to be clustered around the KT boundary [29]. The black dots indicate recent polyploidy events [~1–2 mya in cotton (*Gossypium hirsutum*), <10 mya in potato (*Solanum tuberosum*), ~10–15 mya in soybean (*Glycine max*), ~10 mya in maize (*Zea mays*), and <1 mya in wheat (*Triticum aestivum*)]. Alpha, beta and gamma denote the generally accepted duplication events in *Arabidopsis* [5–7,36] (see main text for details). Modified with permission from [29].

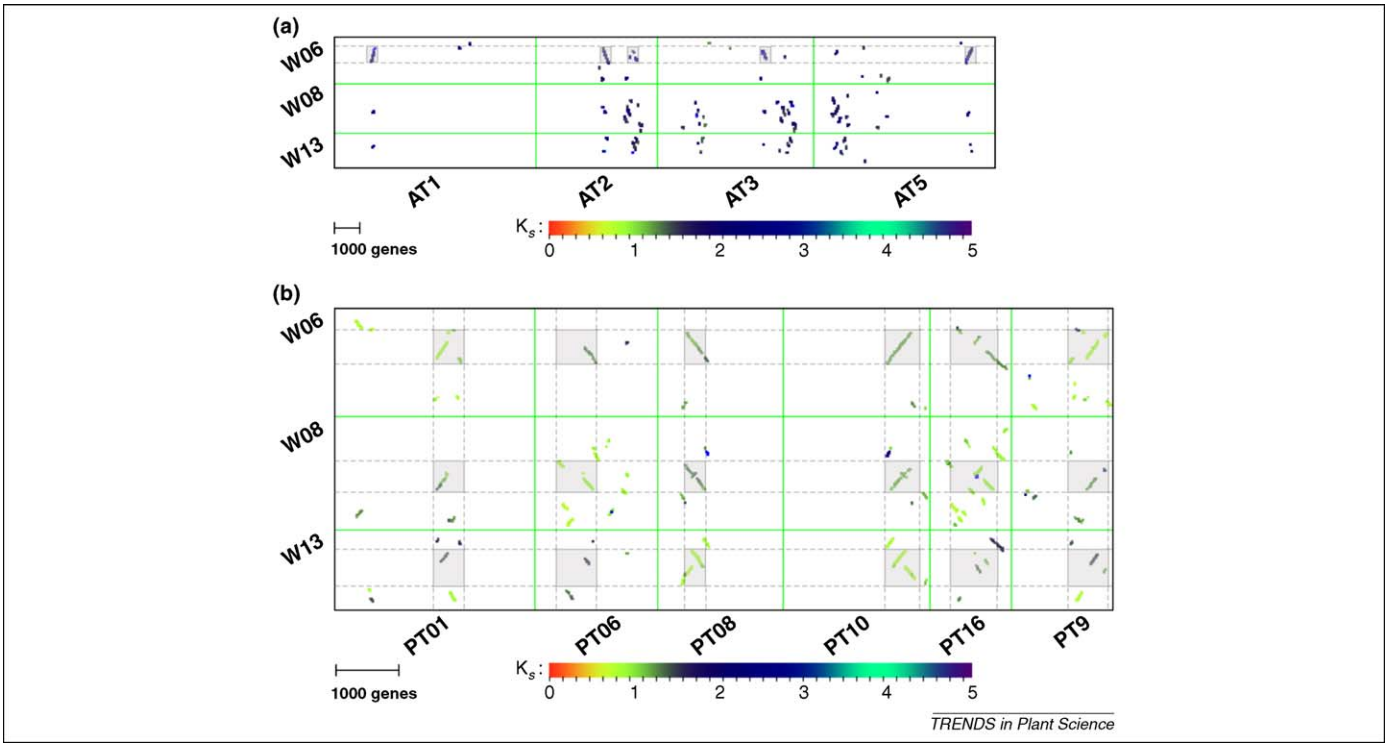

Figure 3. Dot plot showing homology between three grapevine chromosomes (VV) and chromosomes in *Arabidopsis* (AT) **(a)** and poplar (PT) **(b)**. Homologous regions (highlighted in gray) are shown as shorter or longer diagonals and are colored according to their average K_S value, whereby the younger regions are shown in red and yellow and the older in cyan and indigo (see K_S color scheme). The dot plot between grapevine and poplar shows more a greater number of longer blocks, compared with that between grapevine and *Arabidopsis*. Thus, the homology between grapevine and poplar is more obvious than that between grapevine and *Arabidopsis*, probably because of an additional WGD in *Arabidopsis*, higher frequency of rearrangements, and the more slowly evolving genomes of grapevine and poplar [12,17,30]. In several cases, six regions on six chromosomes of poplar show homology with one region in grapevine. Furthermore, four of those appear to be of similar age, whereas two appear younger, suggesting that the hexapolyploidy event is considerably older than the speciation event between grapevine and poplar.

which underwent another duplication to generate a hexaploid, giving rise to the triplicate structure.

A second wave of WGDs?

Apart from the older duplication events shared by most angiosperms, many plant lineages show traces of an additional, independent and more recent genome duplication [22,24,26–28]. Some of the most diverse and species-rich clades, such as Brassicaceae, Poaceae, Fabaceae, Solanaceae and Asteraceae, have all been suggested to have undergone a WGD before their diversification, although the exact timing is yet to be determined [5]. Interestingly, many independent WGDs, such as those in rice, *Medicago truncatula,* tomato, lettuce (*Lactuca sativa*), cotton (*Gossypium hirsutum*), poplar and banana (*Musa spp*.) appear to have occurred ~60–70 mya [28–31] (Figure 2). Recently, it has been suggested that these duplication events are linked to the Cretaceous–Tertiary (KT) extinction event, which is the most recent large-scale mass extinction of plant and animal plant species, including the dinosaurs [29]. However, one should be cautious when linking polyploidy events with adaptations or species diversification, because these are often difficult to test, and require further investigation [32].

Detecting and dating WGDs

Inferring the exact number and timing of WGDs is not straightforward. For instance, based on additional data, the second *Arabidopsis* genome duplication (the *beta* duplication) turned out to be much younger than previously assumed; after the divergence of *Arabidopsis* and papaya lineages, rather than before the divergence of rosids and asterids [7,8,17]. Here, we discuss three commonly used approaches in studying genome duplications; (1) the use of genomic collinearity within and between genomes, (2) K_S plots (from genomic or EST data), and (3) phylogenetic analysis of gene duplications.

Genomic collinearity

Intra-genome comparison of the poplar genome uncovers many homologous segments with a multiplication level of six, which is in agreement with eudicots being ancient hexaploids, to which an additional genome duplication had been added (Figure 1). Inter-genome comparison with a second species, for instance through a dot-plot approach where homologous genes of the two species are visualized as shown in Figure 3, is often useful in observing whether there is an additional WGD in one of the two species. A dot-plot between the genomes of grapevine and poplar shows that one region in grapevine often corresponds to six regions in poplar, and that one region in poplar often corresponds to three regions in grapevine. From this observation, one can conclude that an additional WGD has occurred in the poplar lineage, after its divergence from the grapevine lineage.

The situation for *Arabidopsis* is, at least at first glance, less clear. Intra-genome comparisons show a similar multiplication level as that found in poplar (Figure 1), although sets of homologous segments with a multiplication level of more than six are sometimes uncovered [8]. This would

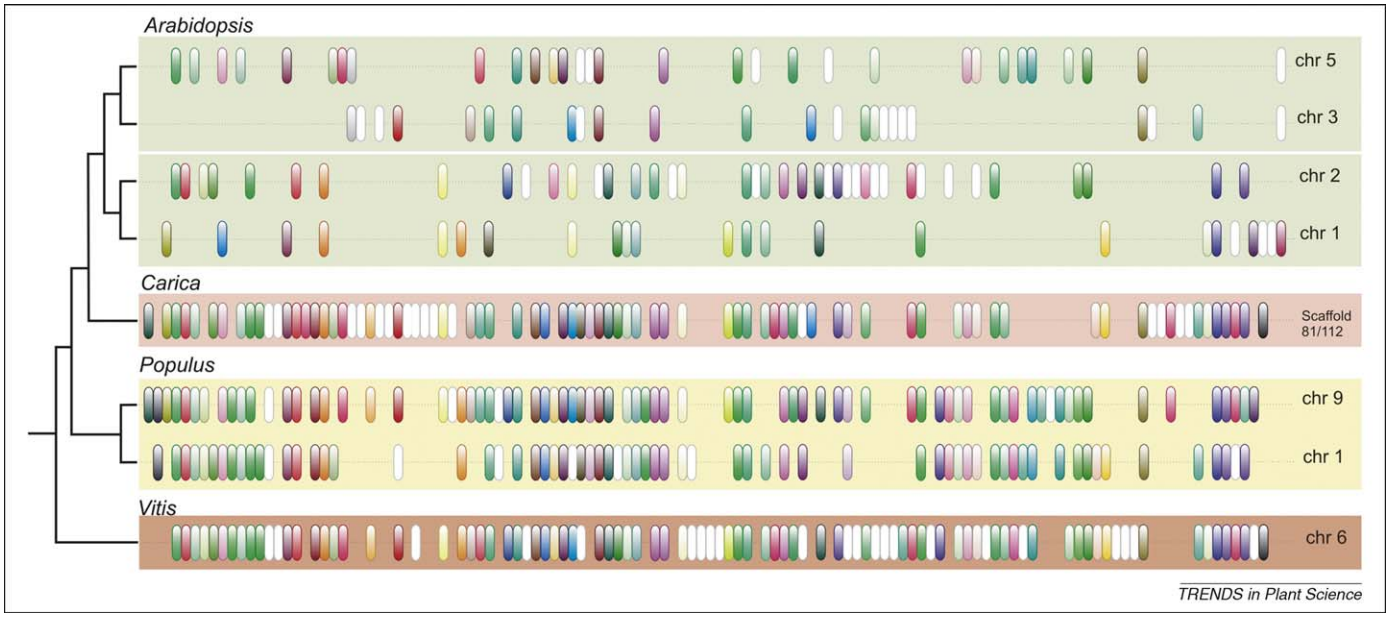

Figure 4. Multiple gene order alignment of four homologous *Arabidopsis*, two poplar, one papaya and one grapevine segment. Genes in the same color belong to the same gene family, whereas genes belonging to families that occur only once are shown in white. Homologous gene families were identified using protein clustering based on BLASTP where all proteins were used as queries against all proteins. All segments were identified as homologous using the i-ADHoRe software tool [61], and using grapevine chromosome 6 as a reference (in case of papaya, two scaffolds were combined to create a larger fragment). As can be observed, few anchors (retained genes in duplicated segments) exist between *Arabidopsis* chromosomes 3 + 5 and 1 + 2, although all four segments show clear homology with the single segments in *papaya* or grapevine. Owing to extensive gene loss and active gene transposition, it is difficult to infer the exact number of duplication events that have occurred in *Arabidopsis* without comparison to a pre-duplicated genome. Multiple gene order alignments can be browsed using the PLAZA comparative genomics platform (http://bioinformatics.psb.ugent.be/plaza/).

argue against two genome duplications in the *Arabidopsis* lineage. However, as previously noted, high frequencies of gene loss (or gene fractionation *sensu* [33]) reduce collinearity resulting in duplicated regions that, in the extreme, no longer share any homologous genes [34]. Nevertheless, by using a top–down approach whereby one simultaneously tries to align structurally similar segments across multiple genomes using transitive homology, highly degenerated homologous segments can often be unveiled [25,35,36]. Figure 4 shows a set of homologous regions in the genomes of grapevine, papaya, poplar and *Arabidopsis* that has been identified by using such a top–down approach. As can be observed, one copy in grapevine and one copy in papaya correspond with two copies in poplar and four in *Arabidopsis,* lending support to one additional duplication in poplar and two in *Arabidopsis* [12,36]. From the top–down approach, it also becomes clear why identifying homologous regions in *Arabidopsis* is difficult when only using intra-genomic analysis; homologous segments in *Arabidopsis* are often highly degenerated. This is also apparent in dot plots where homologous genes of *Arabidopsis* and grapevine, or poplar and grapevine, are compared. Several collinear regions can be identified between grapevine and poplar but not, or with more difficulty, between grapevine and *Arabidopsis* (Figure 3).

Thus, in particular when the species of interest has experienced several rounds of WGDs, has a highly fractionated genome, or experienced severe gene loss, it is recommended to analyze multiple genomes at the same time or use more sophisticated bioinformatics analyses [34–38]. Furthermore, the choice of organism for comparative analysis is also important. For instance, the genome of grapevine has not experienced any additional WGDs after the ancestral hexaploidy (Figure 2) and is evolving slowly, and is therefore thought to still reflect the more ancestral state [12]. Thus, the grapevine genome will be more effective in detecting additional WGDs in other species.

Excessive gene loss and genome rearrangements are the most important causes for breaking up collinearity within and between genomes [39,40]. A recent study suggested that the transposition of genes might also contribute significantly to the loss of collinearity. The level of gene transposition in *Arabidopsis* was determined using papaya as an outgroup species. Starting from collinear regions between both species, the chromosomal positions of *Arabidopsis* genes were scored based on the conservation of homologous neighboring genes in the outgroup species [33]. Although the frequency of transposition varied among different gene families and functional categories, the authors estimated that ~25% of all *Arabidopsis* genes were transposed after the origin of the Brassicales. Therefore, both massive gene loss and gene rearrangements or translocations have been responsible for the highly degenerated patterns of collinearity observed in intra-genome *Arabidopsis* comparisons (Figure 1). Whether the observed rate of gene transposition varies between different angiosperms requires further investigation.

K_S age distributions

Although the identification of inter-genome collinearity using multiple species is probably the most powerful and reliable way to determine the number of WGDs in a given lineage (Figures 3 and 4), such an analysis is often not possible, because it requires a genomic assembly of the species of interest. In many cases, WGDs are inferred by building age distributions of paralogs, where the number of paralogs is plotted against their age, which can be approximated by the number of synonymous substitutions per

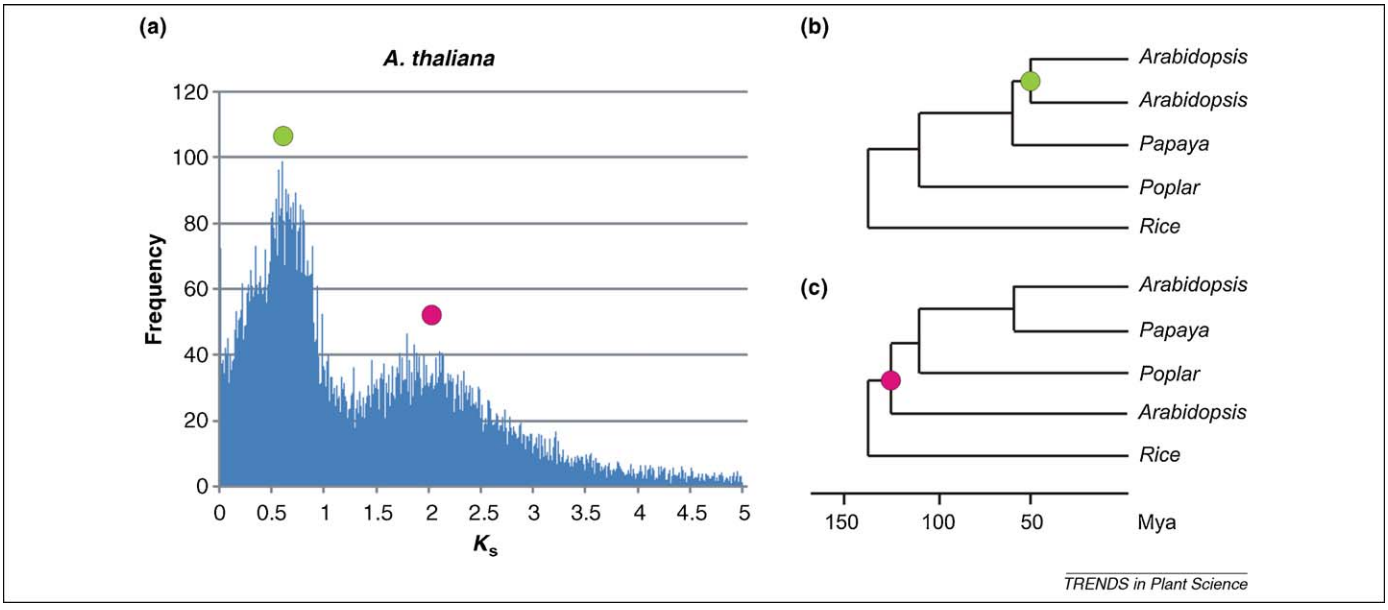

Figure 5. **(a)** Age distribution of the *Arabidopsis* paralogs based on K_S values (number of synonymous substitutions per synonymous site). The peak around $K_S = 0.6$ originates from the youngest genome duplication, while the other, wider peak reflects older genome duplications [17]. **(b, c)** Dating through phylogenetic means. The green dot denotes the youngest large-scale gene duplication event in *Arabidopsis,* while the red dot refers to older duplication events. See text for details.

synonymous site (K_S). A peak in such a distribution indicates a burst of duplications at about the same time, and is often interpreted as a large-scale or WGD event (Figure 5a) [22,26,27,35].

Phylogenetic tree construction

Phylogenetics has often been used to determine the relative (and absolute) timing of WGDs [35]. By mapping the duplication events onto phylogenetic trees that include the paralogs created by a given WGD and orthologous genes from other species, one can determine whether those paralogs were created before or after a given speciation event (Figure 5b,c). For instance, if one wants to determine the timing of a WGD event in *Arabidopsis*, and the dominant topology between two *Arabidopsis* paralogs and the papaya ortholog is ((*Arabidopsis*, *Arabidopsis*), papaya), it is concluded that the WGD event occurred after the divergence of the lineages of *Arabidopsis* and papaya, and is therefore specific to the *Arabidopsis* lineage (Figure 5b). If, by contrast, the most dominant topology is ((*Arabidopsis*, papaya), *Arabidopsis*), then the WGD event is more likely to have occurred before the divergence, and is thus shared between *Arabidopsis* and papaya (Figure 5c). Such an approach was, for instance, used to map the timing of the WGD events in *Arabidopsis*, and it was suggested that the youngest one (*alpha*) is shared with *Brassica*, the second youngest one (*beta*) with tomato (and thus with most eudicots) but not with rice, and the oldest one (*gamma*) with rice but not with gymnosperms [7]. The age of the *beta* event turned out to be highly overestimated, as discussed above (Figure 2)[17].

Also, it was suggested that the most recent WGD in rice was shared with barley (*Hordeum vulgare*) [31] and other cereals [21], and that the WGD in *Medicago* was shared with *Lotus* and soybean (*Glycine max*) [41,42], based on such phylogenetic approaches. A duplication or hybridization event specific to grapevine has been proposed because more topologies of ((grapevine, grapevine), *Arabidopsis*) than ((grapevine, *Arabidopsis*), grapevine) were observed [18]. However, although widely applied, it was recently suggested that the extensive rate variation among species leads to incorrect phylogenetic inferences [17]. Indeed, even if a duplication event was shared between grapevine and *Arabidopsis*, the topology of ((grapevine, grapevine), *Arabidopsis*) can be obtained frequently because the two grapevine sequences will be more similar to each other owing to their slower evolutionary rates, rather than reflecting their true evolutionary relationships. By contrast, regarding *Arabidopsis*, which is thought to have a fast molecular evolution rate [17], even if a duplication event is specific to *Arabidopsis*, a tree where the two *Arabidopsis* paralogs do not cluster together can be obtained frequently owing to their faster evolutionary rate (and in some cases, due to an accelerated evolutionary rate after duplication)[43]. These erroneous phylogenetic inferences that result from extensive rate variation have probably resulted in the incorrect placement of various duplication events and underlie the importance of careful taxon sampling and consideration of possible differences in substitution rates when undertaking phylogenetic studies.

Conclusions and perspectives

Although the number of complete genomes available for analysis is still limited, plant genomes have proved to be particularly exciting. For instance, it was a surprise to discover that the small genome of *Arabidopsis* (~128 Mb) has undergone three paleo-polyploidies, namely one tripling and two doublings, resulting in 12 copies of its ancestral genome [36]. Furthermore, it seems increasingly evident that these genome doublings in flowering plants can be linked to decisive periods in the evolution of the flowering plants, such as the origin and early divergence of the angiosperms, the evolution of flowers or the KT boundary [4,29,32,44].

As we have discussed here, WGDs are an inextricable part of the evolution of angiosperms and the recent

increase in the number of genomic sequences, together with the improvement in the methods to determine the number and timing of WGDs has already resulted in a better understanding of the impact of WGDs on the evolution of angiosperms [4,32,45,46]. The rapid increase in the number of genomes being sequenced will continue to enable researchers to determine more accurately the number and timing of WGDs in different plant lineages, which will in turn enable issues such as the extent to which WGDs have facilitated avoiding extinctions, adaptation and diversification, to be better addressed [44]. However, for now, it is difficult to know whether we could have marveled at the diversity of the Amazon forest, the splendor of the spring flowers in Namakwaland, or the many colors and scents in flower shops, if those large-scale gene duplications in flowering plants had not occurred, just as we might not know whether complex vertebrates, such as humans, would have evolved to enjoy all that beauty if they had not doubled their own genome twice >500 mya [44,56].

Acknowledgements

We thank the Bioinformatics and Evolutionary Genomics team, as well as Riccardo Velasco and Francesco Salamini for discussions. Steven Maere is acknowledged for technical help. S.P. is indebted to the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT). K.V. is a postdoctoral fellow of the Fund for Scientific Research Flanders (FWO). This work is supported by the EU [EU-FP6 Food Safety and Quality: FOOD-CT-2006-016214 (EU-SOL)] and by the Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet). We also thank the referees for their thoughtful suggestions and comments, which have greatly improved this review.

References

1 Friis, E.M. *et al.* (2006) Cretaceous angiosperm flowers: innovation and evolution in plant reproduction. *Palaeogeogr. Palaeocl.* 232, 251–293
2 Moore, M.J. *et al.* (2007) Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proc. Natl. Acad. Sci. U. S. A.* 104, 19363–19368
3 Crepet, W.L. and Niklas, K.J. (2009) Darwin's second "abominable mystery": why are there so many angiosperm species? *Amer. J. Bot.* 96, 366–381
4 De Bodt, S. *et al.* (2005) Genome duplication and the origin of angiosperms. *Trends Ecol. Evol.* 20, 591–597
5 Soltis, D.E. *et al.* (2009) Polyploidy and angiosperm diversification. *Am. J. Bot.* 96, 336–348
6 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
7 Bowers, J.E. *et al.* (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422, 433–438
8 Simillion, C. *et al.* (2002) The hidden duplication past of *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U. S. A.* 99, 13627–13632
9 Zahn, L.M. *et al.* (2005) The evolution of the SEPALLATA subfamily of MADS-box genes: a preangiosperm origin with multiple duplications throughout angiosperm history. *Genetics* 169, 2209–2223
10 Maere, S. *et al.* (2005) Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* 102, 5454–5459
11 Veron, A.S. *et al.* (2007) Evidence of interaction network evolution by whole-genome duplications: a case study in MADS-box proteins. *Mol. Biol. Evol.* 24, 670–678
12 Jaillon, O. *et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449, 463–467
13 Ming, R. *et al.* (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* 452, 991–996
14 Wikström, N. *et al.* (2001) Evolution of the angiosperms: calibrating the family tree. *Proc. R Soc. Lond. B* 268, 2211–2220
15 Jansen, R.K. *et al.* (2006) Phylogenetic analyses of *Vitis* (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. *BMC Evol. Biol.* 6, 32
16 Jansen, R.K. *et al.* (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proc. Natl. Acad. Sci. U. S. A.* 104, 19369–19374
17 Tang, H. *et al.* (2008) Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* 18, 1944–1954
18 Velasco, R. *et al.* (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE* 2, e1326
19 Paterson, A.H. *et al.* (2005) Ancient duplication of cereal genomes. *New Phytol.* 165, 658–661
20 Zhang, Y. *et al.* (2005) Two ancient rounds of polyploidy in rice genome. *J. Zhejiang Univ. Sci.* 6B, 87–90
21 Vandepoele, K. *et al.* (2003) Evidence that rice and other cereals are ancient aneuploids. *Plant Cell* 15, 2192–2202
22 Cui, L. *et al.* (2006) Widespread genome duplications throughout the history of flowering plants. *Genome Res.* 16, 738–749
23 Bell, C.D. *et al.* (2005) The age of the angiosperms: a molecular timescale without a clock. *Evolution* 59, 1245–1258
24 Barker, M.S. *et al.* (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
25 Lyons, E. *et al.* (2008) The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates the rosids. *Tropical Plant Biol.* 1, 181–190
26 Blanc, G. and Wolfe, K.H. (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* 16, 1667–1678
27 Schlueter, J.A. *et al.* (2004) Mining EST databases to resolve evolutionary events in major crop species. *Genome* 47, 868–876
28 Lescot, M. *et al.* (2008) Insights into the *Musa* genome: syntenic relationships to rice and between Musa species. *BMC Genomics* 9, 58
29 Fawcett, J.A. *et al.* (2009) Plants with double genomes might have had a better chance to survive the Cretaceous- Tertiary extinction event. *Proc. Natl. Acad. Sci. U. S. A.* 106, 5737–5742
30 Tuskan, G.A. *et al.* (2006) The genome of black cottonwood. *Populus trichocarpa (Torr. & Gray). Science* 313, 1596–1604
31 Paterson, A.H. *et al.* (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl. Acad. Sci. U. S. A.* 101, 9903–9908
32 Soltis, D.E. and Burleigh, J.G. (2009) Surviving the K-T mass extinction: new perspectives of polyploidization in angiosperms. *Proc. Natl. Acad. Sci. U. S. A.* 106, 5455–5456
33 Freeling, M. *et al.* (2008) Many or most genes in *Arabidopsis* transposed after the origin of the order Brassicales. *Genome Res.* 18, 1924–1937
34 Vandepoele, K. *et al.* (2002) Detecting the undetectable: uncovering duplicated segments in *Arabidopsis* by comparison with rice. *Trends Genet.* 18, 606–608
35 Van de Peer, Y. (2004) Computational approaches to unveiling ancient genome duplications. *Nat. Rev. Genet.* 5, 752–763
36 Tang, H. *et al.* (2008) Synteny and collinearity in plant genomes. *Science* 320, 486–488
37 Simillion, C. *et al.* (2004) Building genomic profiles for uncovering segmental homology in the twilight zone. *Genome Res.* 14, 1095–1106
38 Soltis, P.S. and Soltis, D.E. (2009) The role of hybridization in plant speciation. *Annu. Rev. Plant Biol.* 60, 561–588
39 Thomas, B.C. *et al.* (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
40 Zheng, C. *et al.* (2009) Gene loss under neighborhood selection following whole genome duplication and the reconstruction of the ancestral *Populus* genome. *J. Bioinform. Comput. Biol.* 7, 499–520
41 Cannon, S.B. *et al.* (2006) Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* genomes. *Proc. Natl. Acad. Sci. U. S. A.* 103, 14959–14964
42 Pfeil, B.E. *et al.* (2005) Placing paleopolyploidy in relation to taxon divergence: a phylogenetic analysis in legumes using 39 gene families. *Syst. Biol.* 54, 441–454

43 Van de Peer, Y. *et al.* (2002) Dealing with saturation at the amino acid level: a case study based on anciently duplicated zebrafish genes. *Gene* 295, 205–211
44 Van de Peer, Y. *et al.* (2009) The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.* 10, 725–732
45 Bowman, J.L. *et al.* (2007) Green genes-comparative genomics of the green branch of life. *Cell* 129, 229–234
46 Soltis, D.E. *et al.* (2008) Origin and early evolution of angiosperms. *Ann. NY Acad. Sci.* 1133, 3–25
47 Swarbreck, D. *et al.* (2008) The *Arabidopsis* Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.* 36, D1009–1014
48 Sato, S. *et al.* (2008) Genome Structure of the Legume. *Lotus japonicus. DNA Res.* 15, 227–239
49 Sterck, L. *et al.* (2007) How many genes are there in plants (. . . and why are they there)? *Curr. Opin. Plant Biol.* 10, 199–203
50 Tanaka, T. *et al.* (2008) The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Res.* 36, D1028–1033
51 Vandepoele, K. and Van de Peer, Y. (2005) Exploring the plant transcriptome through phylogenetic profiling. *Plant Phys.* 137, 31–42
52 Seoighe, C. and Gehring, C. (2004) Genome duplication led to highly selective expansion of the *Arabidopsis thaliana* proteome. *Trends Genet.* 20, 461–464
53 Blanc, G. and Wolfe, K.H. (2004) Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell* 16, 1679–1691
54 Chapman, B.A. *et al.* (2006) Buffering of crucial functions by paleologous duplicated genes may contribute cyclicality to angiosperm genome duplication. *Proc. Natl. Acad. Sci. U. S. A.* 103, 2730–2735
55 Schranz, M.E. and Mitchell-Olds, T. (2006) Independent ancient polyploidy events in the sister families Brassicaceae and Cleomaceae. *Plant Cell* 18, 1152–1165
56 Blomme, T. *et al.* (2006) The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol.* 7, R43
57 Davis, J.C. and Petrov, D.A. (2005) Do disparate mechanisms of duplication add similar genes to the genome? *Trends Genet.* 21, 548–551
58 Freeling, M. and Thomas, B.C. (2006) Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res.* 16, 805–814
59 Papp, B. *et al.* (2003) Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424, 194–197
60 Krylov, D.M. *et al.* (2003) Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res.* 13, 2229–2235
61 Simillion, C. *et al.* (2008) i-ADHoRe 2.0: an improved tool to detect degenerated genomic homology using genomic profiles. *Bioinformatics* 24, 127–128