Gamma Paleohexaploidy in the Stem Lineage of Core Eudicots: Significance for MADS-Box Gene and Species Diversification

Dries Vekemans,[†,1] Sebastian Proost,[†,2,3] Kevin Vanneste,[2,3] Heleen Coenen,[1] Tom Viaene,[1] Philip Ruelens,[1] Steven Maere,[2,3] Yves Van de Peer,[*,2,3] and Koen Geuten[*,1]

[1]Department of Biology, KULeuven (University of Leuven), Leuven, Belgium

[2]Department of Plant Systems Biology, VIB, Ghent, Belgium

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

[†]These authors contributed equally to this work.

***Corresponding author:** yves.vandepeer@psb.vib-ugent.be; koen.geuten@bio.kuleuven.be.

Associate editor: Michael Purugganan

Data deposition: Protein sequences reported in this article have been deposited in Genbank (accession nos. JX266529–JX266563).

Abstract

Comparative genome biology has unveiled the polyploid origin of all angiosperms and the role of recurrent polyploidization in the amplification of gene families and the structuring of genomes. Which species share certain ancient polyploidy events, and which do not, is ill defined because of the limited number of sequenced genomes and transcriptomes and their uneven phylogenetic distribution. Previously, it has been suggested that most, but probably not all, of the eudicots have shared an ancient hexaploidy event, referred to as the gamma triplication. In this study, detailed phylogenies of subfamilies of MADS-box genes suggest that the gamma triplication has occurred before the divergence of Gunnerales but after the divergence of Buxales and Trochodendrales. Large-scale phylogenetic and K_S-based approaches on the inflorescence transcriptomes of *Gunnera manicata* (Gunnerales) and *Pachysandra terminalis* (Buxales) provide further support for this placement, enabling us to position the gamma triplication in the stem lineage of the core eudicots. This triplication likely initiated the functional diversification of key regulators of reproductive development in the core eudicots, comprising 75% of flowering plants. Although it is possible that the gamma event triggered early core eudicot diversification, our dating estimates suggest that the event occurred early in the stem lineage, well before the rapid speciation of the earliest core eudicot lineages. The evolutionary significance of this paleopolyploidy event may thus rather lie in establishing a species lineage that was resilient to extinction, but with the genomic potential for later diversification. We consider that the traits generated from this potential characterize extant core eudicots both chemically and morphologically.

Key words: hexaploidy, Gunnerales, Buxales, whole-genome duplication, diversification, gamma.

Introduction

Ancient polyploidy events have long been thought of to play a significant role in shaping biodiversity, but only in the last decade, clear evidence for paleopolyploidy was found through comparative genomics (Wolfe and Shields 1997; Simillion et al. 2002; Taylor et al. 2003; Kellis et al. 2004; Van de Peer 2004). As gene duplication is considered to provide genetic material for the evolution of novel traits and functions, whole-genome duplications (WGDs) might be events with a transformative evolutionary potential. Whether and how the polyploid state results in a competitive advantage over diploids remains difficult to assess (Soltis and Soltis 2000; Otto 2007; Fawcett and Van de Peer 2010). In plants, polyploidy occurs frequently and an estimated 35% of species have become polyploid after the origin of their genus (Wood et al. 2009). Such recent polyploids most often represent evolutionary dead ends (Otto and Whitton 2000; Mayrose et al. 2011). Therefore ancient polyploids appear to be exceptions that possibly survived because of intrinsic differences compared with most recent polyploids or because of tailored ancient ecological opportunities.

Some short- and long-term effects of WGD have been assessed or predicted, but many remain to be resolved (Otto 2007; Van de Peer et al. 2009). Immediately after polyploidization, the genome is expected to structurally reorganize with changes in gene expression (Liu and Wendel 2003; Adams and Wendel 2005). The polyploid is possibly more resilient to environmental stress and may be able to establish itself in different ecological habitats because of affected traits such as drought resistance, flowering time, plant-pollinator interactions, etc. (Levin 1983; Ramsey 2011). Establishing itself in a population probably is difficult, as fitness is often reduced and crossing with diploids is unlikely to result in viable offspring (Ramsey and Schemske 1998). Although negative mutations are better masked in the polyploid population, this effect is only transitory (Otto and Whitton 2000; Otto 2007). The expected longer-term advantage of ancient polyploids is to come from retained duplicated genes that acquire

a new function or partition functions (Ohno 1970; Force et al. 1999; Freeling 2009). Such duplicates are most likely to function in dosage balanced regulatory processes like transcriptional regulation and signal transduction (Blanc and Wolfe 2004; Maere et al. 2005; Freeling and Thomas 2006; Hakes et al. 2007; Freeling 2009; Birchler and Veitia 2010). Ultimately, if the polyploid lineage survives to generate novel species, it is not clear whether the descendant species groups should be expected to be more diverse or less diverse (Vamosi and Dickinson 2006; Soltis et al. 2009; Mayrose et al. 2011). Whereas the identification of ancient WGD events has fueled the idea of a causal relation between paleopolyploidy and successful taxonomic groups, the nature of this relation is not well defined (Otto 2007; Soltis et al. 2009; Van de Peer et al. 2009; Jiao et al. 2011, 2012). Possibly, paleopolyploids trigger speciation, but maybe they rather carry the genomic potential for later diversification.

To evaluate the evolutionary significance of ancient WGDs, it is necessary to date these events accurately and relative to other events, such as species diversifications and geological occurrences. Similar to how species phylogenies are used to understand character evolution, accurate phylogenetic placement allows to improve our inferences of the ancestral morphology and ecology of the paleopolyploid and opens avenues to better understand the relevance of these events for taxonomic diversity. Current evidence indicates that the ancestor of all extant seed plants was polyploid and that all flowering plants share a second paleopolyploid ancestor after the divergence of gymnosperms, coined epsilon (Jiao et al. 2011). Later, in flowering plant evolution, in association with the divergence of core eudicots, a triplication coined gamma occurred which was recently estimated at 117 Ma (Jaillon et al. 2007; Jiao et al. 2012). Recently, Jiao et al. (2012) provided strong evidence that the gamma polyploidy event occurred early in eudicot evolution. However, although the "basal" eudicot Ranunculales and Proteales transcriptomes were included in their analyses, Buxales, Trochodendrales and the earliest diverging core eudicot lineages were not included. Therefore, the precise positioning of this event in the species phylogeny remains unresolved. Apart from the ancient WGDs that occurred in the ancestors of seed and flowering plants (Jiao et al. 2011) and the gamma hexaploidization, there seems to have been a more recent wave of WGDs as well. It has been suggested that independent WGDs have occurred in many lineages close to the Cretaceous-Tertiary mass extinction event 65 Ma (Fawcett et al. 2009). This correlation in time raises the question whether polyploids may resist extinction better than diploids (Crow and Wagner 2006). The survival potential of polyploids could have allowed their descendant lineages to occupy empty niches after an extinction event during vegetation recovery. Alternatively, paleopolyploid lineages may have had an intrinsic capacity to diversify, which has not been realized (yet) in more recent polyploids (Soltis et al. 2009; Mayrose et al. 2011).

Similar to the duplication history of HOX genes in vertebrates, duplications in plant MADS-box transcription factors have been studied to understand the origins and evolution of plant developmental mechanisms (e.g., Theissen and Melzer 2007; Geuten and Irish 2010). The authors, and others observed that early in eudicot evolution, several subfamilies of MADS-box genes have duplicated or triplicated (Kramer et al. 1998; Litt and Irish 2003; Zahn et al. 2005, 2006; Kramer et al. 2006; Shan et al. 2007; Viaene et al. 2010). The phylogenetic placement of the gamma event by Jiao et al. (2012) already makes it plausible that these MADS-box gene duplications may derive from the gamma triplication, given the approximate correspondence of these events in time. But demonstrating this exactly is important, for example, to evaluate the evolution of the MADS–domain interaction network after WGDs (Veron et al. 2007).

Core eudicots have been defined as including all *Superasteridae* (comprising Berberidopsidales, Santalales, Caryophyllales, *Asteridae,* and Dilleniaceae) and *Superrosidae* (comprising *Rosidae* (including Vitaceae) and Saxifragales). Sister to all other core eudicots is the small order Gunnerales (fig. 1; Soltis et al. 2011). Molecular dating studies have inferred that the earliest lineages of crown group core eudicots diverged in a narrow window of time, presumably in less than one million year. Yet, the stem lineage of core eudicots did not generate extant species lineages for a longer time, possibly for >7 My (Magallón and Castillo 2009; Moore et al. 2010). This presents the question what may have triggered this sudden diversification, and whether it could be directly related to the occurrence of a WGD at the base of the core eudicot crown group. The core eudicots as a group are characterized by a number of derived traits, with the production of ellagic and gallic acids, the canalization of floral organ number and a clear separation of sepal and petal identity, with petals probably being derived from bracts (Stevens 2001; Soltis et al. 2005, De Craene 2007). It has been suggested that MADS-box gene duplications were involved in the origin of these reproductive traits and in the diversification of core eudicot reproductive morphology (Irish and Litt 2005; Irish 2006).

In this study, we provide detailed phylogenies of MADS-box genes families that place the gamma triplication at the precise origin of core eudicots, before Gunnerales and after Buxales branch off in the species phylogeny. To provide unambigous and large-scale support for this placement of gamma, we also performed genome-wide phylogenetic analyses in combination with K_S-based dating and age distribution mixture modeling on the transcriptomes of *G. manicata* (Gunneraceae, Gunnerales) and *P. terminalis* (Buxaceae, Buxales). The accurate phylogenetic placement of the gamma triplication allows us to make a first evaluation of the potential of this genomic event for the species lineages that derived from it.

Materials and Methods

RNA Isolation and Cloning of MADS-box Genes

Floral buds from *G. manicata* (Gunneraceae), *P. terminalis* (Buxaceae), *Nyssa sylvatica* (Nyssaceae), *Actinidia chinensis* (Actinidiaceae), *Heliamphora minor* (Sarraceniaceae), *Jacquinia aurantiaca* (Theophrastaceae), *Styrax japonicus, Halesia*

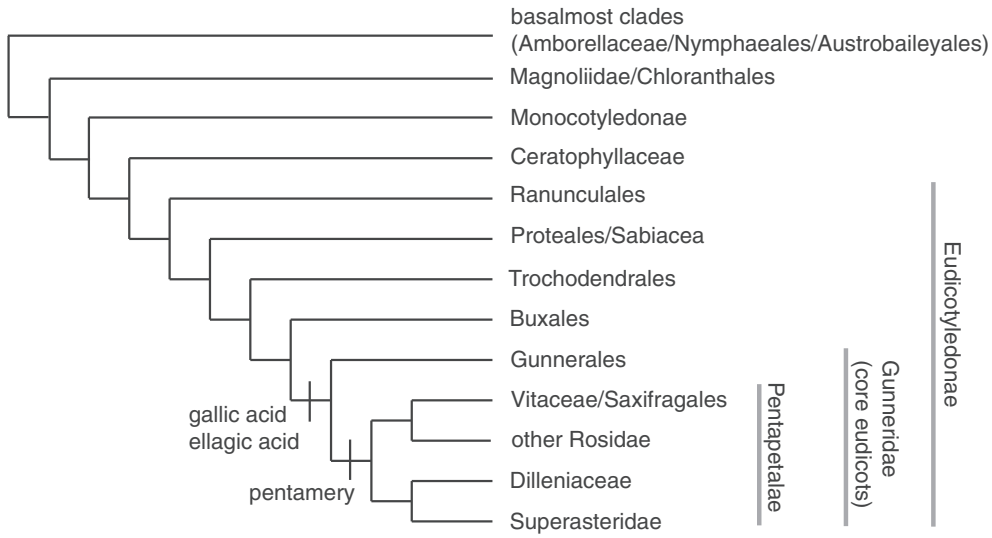


Fig. 1. Phylogeny of early diverging eudicot and core eudicot lineages. Core eudicots include Gunnerales but not Buxales (Soltis et al. 2003, Angiosperm Phylogeny Group 2009). The phylogeny is redrawn from (Soltis et al. 2011). Note that a different relative branching order has been obtained for the phylogenetic placement of Trochodendrales/Buxales, Vitaceae/Saxifragales, and Dilleniaceae in other recent analyses (Moore et al. 2011).

diptera (Styracaceae), *Clethra tomentosa* (Clethraceae), *Erica hiemales* (Ericaceae), *Camellia japonica*, *Stewartia pseudocamellia* (Theaceae), and *Ipomopsis aggregata* (Polemoniaceae) were frozen in liquid nitrogen and stored at −80°C. Total RNA was isolated using the Invisorb Spin Plant RNA MINI Kit (Invitek, Berlin, DE) or Trizol (Invitrogen, Carlsbad, USA). The mRNA was reverse transcribed into cDNA using *AMV reverse transcriptase* (Promega, Madison, USA) with an oligo-dT primer (Kramer et al. 1998). Except for *G. manicata* and *P. terminalis*, *AGL234-like* sequences for all collected species were amplified using a degenerate forward primer RQVT (5′-CGRCARGTGACSTTCTSCAARCG-3′) and a PCR-program taken from the literature (Kramer et al. 1998; Winter et al. 1999). All PCR amplifications were carried out using *Taq* DNA Polymerase (Invitrogen, Carlsbad, USA). PCR products were gel-purified using the Nucleospin extract 2 kit (Macherey-Nagel, Düren, DE) and cloned into the pGEM-T vector (Promega, Madison, USA). After transformation, between 50 and 100 white clones were checked for inserts in a PCR reaction using the same primers and program. Plasmid DNA for selected clones was extracted with a Nucleospin Plasmid kit (Macherey-Nagel, Düren, DE) or a PureYield Plasmid Miniprep System kit (Promega, Madison, USA). The plasmids were sent for sequencing (MacroGen Inc., Seoul, KP).

Transcriptome Sequencing and Assembly

The RNA samples for next-generation sequencing of *G. manicata* and *P. terminalis* inflorescence tissue were isolated using the same method, but were subsequently DNase treated using TURBO DNA-free (Ambion, Austin, USA). We tested the RNA quality using an Agilent 2100 BioAnalyzer (~20 μg at a concentration of $\geq$400 ng/μl and a RIN-value of 7). RNA pair-end sequencing was perfomed by the Beijing Genomics Institute (BGI) on an Illumina HiSeq 2000 instrument, after enriching the mRNA (using oligo(dT) and fragmenting the mRNA (200 nt–700 nt). Low quality reads and reads containing adapter sequences were removed, the remaining reads were assembled using SOAPdenovo (Li et al. 2010). Sequence assembly was performed by BGI.

Phylogenetic Analyses of MADS-Box Genes

To identify the exact phylogenetic position of the gamma triplication, we assembled data matrices for six different MADS-box gene lineages (*AG*, *DEF*, *AGL2-3-4*, *SQUA*, *SOC1/TM3*, and *AGL6*) (supplementary table S1, Supplementary Material online, for a list of species used in phylogenetic and dating analyses with abbreviations) with selected representatives from basal angiosperms, monocots, basal eudicots, and core eudicots. Starting from a MUSCLE (Edgar 2004) alignment with default parameters, the alignment was manually improved using MacClade4 (Maddison 2000). Modeltest 3.06 (Posada and Crandall 1998) selected the GTR+I+G substitution model under the Akaike Information Criterion for each data set. We used PhyML for maximum likelihood inference (Guindon and Gascuel 2003) and node support was evaluated through bootstrap analyses with 100 replicates. The most likely tree was used to plot bootstrap values (>50). The resulting trees can be found in supplementary figures S1–S6, Supplementary Material online.

Divergence Time Estimates

To estimate divergence times for the different MADS-box gene lineages, we first rearranged the topology of our maximum-likelihood trees manually in MacClade4 to correspond to the species phylogeny presented in Soltis et al. 2011. Next, we imported these trees into RAxML 7.04 to re-estimate branch lengths under the GTR + I + G model. The obtained trees were then used as input files in r8s (Sanderson 2002). We applied the semiparamatric penalized likelihood approach using a truncated Newton method with bound

constraints, as implemented in r8s. Under the assumption of a relaxed molecular clock, we used a cross-validation procedure to determine the optimal smoothing parameter for each analysis.

With the exception of the eudicot and angiosperm crown group, fossil ages were always treated as minimal ages. We used two approaches to calibrate divergence times. In a first approach we set the angiosperm crown group to a maximal age of 350 My to correspond to the age of the most recent common ancestor (MRCA) of extant seed plants (Rothwell and Scheckler 1988). As the estimated clade ages depended mainly on the number of the included fossil constraints, we chose in a second approach to apply a maximum age of 130 My for the angiosperm crown group (Hughes and McDougall 1990; Hughes et al. 1991; Brenner 1996) and to fix the stem lineage of Gunnerales to 116.74 My following (Magallón and Castillo 2009). All fossil age constraints are summarized in supplementary table S2, Supplementary Material online, and were taken from Magallón and Castillo (2009).

Construction of Data Sets

The genomes for rice (*Oryza Sativa*) (Goff et al. 2002), grape (*Vitis vinifera*) (Jaillon et al. 2007), *Arabidopsis thaliana* (Arabidopsis Genome Initiative 2000), apple (*Malus domestica*) (Velasco et al. 2010), and poplar (*Populus trichocarpa*) (Tuskan et al. 2006) were obtained from PLAZA 2.5 (Van Bel et al. 2012) (http://bioinformatics.psb.ugent.be/plaza/). Assembled ESTs for sunflower (*Helianthus annuus*) and *Aquilegia formosa* x *A. pubescens* were downloaded from TIGR Plant Transcript Assemblies (Childs et al. 2007) (http://plantta.jcvi.org/).

Open Reading Frame Detection

For the *H. annuus* and *A. formosa* x *A. pubescens* transcripts, as well as the *Gunnera* and *Pachysandra* transcripts, the open reading frames were predicted using FrameDP version 1.0.3 (Gouzy et al. 2009). As a reference database, all protein-coding genes present in PLAZA 2.5 (Van Bel et al. 2012) were provided. FrameDP was configured to run with blastall 2.2.17 (settings used; Expectation value: 1 e-3, Open Gap Penalty: 9, Gap Extention Penalty: 2 and retaining only the 100 best hits), whereas the GC3 split training with three iterations was used. Other parameters were left at their default values.

Gene Family Construction

Gene families were constructed using an in-house PLAZA pipeline (Proost et al. 2009) after parsing all genomes and transcriptomes to the correct format and uploading them into a relational database. First, all pairwise similarities between proteins were calculated using BLASTP (Altschul et al. 1997) and homologous genes were clustered into gene families using TRIBE-MCL (Enright et al. 2002). In addition to gene families, orthologous groups were generated using OrthoMCL (Li et al. 2003).

Multiple Sequence Alignments and Phylogenetic Tree Construction

For each gene family and orthologous group amino acid sequences of genes from *O. sativa*, *P. terminales*, *G. manicata*, *V. vinifera*, *A. thaliana*, and *P. trichocarpa* were aligned using MUSCLE (Edgar 2004). Highly diverged and partial genes were removed from the alignment if they contained gaps in $>50\%$ of the alignment or twice the number of gaps than the average. Next, dubious positions (positions with gaps in $>10\%$ of the sequences) were removed from the original alignment (as described in Proost et al. 2009). Starting from these stripped alignments, PhyML (using the default settings for protein sequences) was used to generate phylogenetic trees (Guindon et al. 2010) on which automatic tree reconciliation was performed using NOTUNG 2.6 (Chen et al. 2000)

Placing the Gamma Event on Reconciled Gene Trees

For each tree, all nodes were traversed using a custom script to detect duplication events (with sufficient bootstrap support) in the Eudicotyledoneae and core eudicots. For each valid node, the duplication consistency score was calculated (Vilella et al. 2009). Only nodes with a duplication consistency of ≥ 0.5 were considered to avoid including erroneous duplication nodes. The number of trees containing a valid duplication node in the MRCA of the core eudicots + Buxales were counted as were the number of trees supporting a duplication in the MRCA of the core eudicots (after the speciation of the Buxales). For the latter, *V. vinifera* genes derived from the gamma event, as detected using i-ADHoRe (Proost et al. 2012), were cross-compared with the trees to confirm that the detected duplication node was in fact indicative for the triplication.

K_S-Based Dating and Mixture Modeling

For each gene family, all possible pairs of genes from the same species (*Vitis*, *Gunnera*, and *Pachysandra*) were generated along with all combinations of grape and *Gunnera* genes. ClustalW (Thompson et al. 1994) was used to generate codon alignments for each pairwise comparison. K_S estimates were obtained through Maximum Likelihood Estimation (MLE) using the CODEML program (Goldman and Yang 1994) of the PAML package (v4.4 c) (Yang 2007). Codon frequencies were calculated based on the average nucleotide frequencies at the three codon positions (F3 $\times$ 4), and a constant K_N/K_S (reflecting selection pressure) was assumed for every pairwise comparison (codon model 0). For each pairwise comparison, K_S estimation was repeated three times to avoid suboptimal estimates because of MLE entrapment in local maxima. Only K_S estimates <2 were considered in the construction of empirical age distributions. Gene families were subdivided into subfamilies for which K_S estimates between genes did not exceed a value of 2. To correct for the redundancy of K_S values [a gene family of n members produces $n(n-1)/2$ pairwise K_S estimates for $n-1$ retained duplication events], an average linkage clustering approach was used as described in Maere et al. (2005).

We employed a mixture modeling strategy to identify WGDs in the K_S-based age distributions. WGDs result in a sudden burst of new gene duplicates concentrated in time, recognizable as spikes superimposed on an exponential decay distribution of small-scale duplications (SSDs) in the age distribution (Blanc and Wolfe 2004). The compounded distribution is hence expected to consist of a mixture of log-transformed SSD exponentials and WGD Gaussians (Schlueter et al. 2004; Cui et al. 2006). We used the EMMIX software (McLachlan et al. 1999) to fit a mixture model of Gaussian distributions to the log-transformed K_S distributions. The mixture model identifies the number of normal distributions and their positions that best explain the empirical age distributions. All observations <0.1 K_S were removed to avoid the incorporation of allelic and/or splice variants often encountered in transcriptome data sets (Baker 2012), and to prevent the fitting of a component to infinity (Schlueter et al. 2004). We fitted one to five normal components per mixture model to the data, using 1,000 random and 100 k-mean starts. The Bayesian Information Criterion (BIC) (Schwarz 1978) was used to select the best number of components for the mixture model, because it strongly penalizes increases in the number of model parameters to avoid overfitting of components. The mean and variance of each component were back-transformed to the original scale for plotting and interpretation.

For three out of four age distributions (*Pachysandra*, *Vitis*, and the *Gunnera*–*Vitis* orthologs), BIC selected the maximum number of allowed components, hinting that the usage of the BIC model selection criterion still results in overfitting of components (Naik et al. 2007). Following Barker et al. (2008), we therefore employed the SiZer software (Chaudhuri and Marron 1999) to help identify significant features ($\alpha = 0.05$) in the age distributions. This software uses changes in the first derivative of a range of kernel density estimates with different smoothing bandwidths to distinguish peaks in the distribution that represent true features from those that represent noise. Components of the mixture model corresponding to significant features as identified by SiZer were then interpreted in light of the paleopolyploid history (significant features identified by SiZer corresponding to components fitting the SSD background were excluded). Additionally, we also extended the maximum number of fitted components from five to ten. For two out of three age distributions (*Pachysandra* and *Vitis*), the BIC criterion identified a higher number of components (six and seven, respectively), but this did not influence the number of significant (WGD) peaks identified by SiZer (supplementary table S3, Supplementary Material online).

Detection of Collinearity in Vitis

Using i-ADHoRe 3.0 (Proost et al. 2012) collinear regions in the *Vitis* genome were detected using only the *Vitis* genome and our gene families as input. Settings used were: alignment_method gg2, gap_size 30, tandem_gap 30, cluster_gap 35, q_value 0.85, prob_cutoff 0.01, multiple_hypothesis_correction FDR, anchor_ points 5, and level_2_only false.

Results

Gamma-Derived Duplications in MADS-Box Subfamilies Trace Back to the Stem Lineage of Core Eudicots

We reasoned that by demonstrating orthology of WGD-derived duplicates to genes sampled throughout the species phylogeny, it is possible to putatively place the WGD event in the species phylogeny. Duplications in several MADS-box gene subfamilies are known to have occurred close to the origin of core eudicots (Kramer et al. 1998; Litt and Irish 2003; Zahn et al. 2005, 2006; Kramer et al. 2006; Wu and Su 2007; Shan et al. 2007; Viaene et al. 2010). In these studies, a different species sampling was used each time and some informative core eudicot taxa were not consistently sampled. More importantly, it was not shown that the duplicate lineages derive from the gamma triplication.

To establish this unambiguously, we constructed detailed phylogenies for the *AGAMOUS* (*AG*), *DEFICIENS* (*DEF*), *SEPALLATA1/2/4* (*SEP1/2/4*), *SQUAMOSA* (*SQUA*), *SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1* (*SOC1*), and *AGAMOUS-LIKE6* (*AGL6*) subfamilies (fig. 2 and supplementary figs. S1–S6, Supplementary Material online). *Vitis vinifera* members of each of these subfamilies have been mapped to syntenic regions in the *Vitis* genome that are duplicated or triplicated and derive from the gamma triplication (Jaillon et al. 2007; Díaz-Riquelme et al. 2009). For the *AG*, *SQUA*, and *SOC1* subfamilies, multiple *Vitis* gene family members were found as conserved duplicates in collinear regions using i-ADHoRe 3.0 (Proost et al. 2012), which supports that these duplicates were derived from the gamma WGD (fig. 2*B*).

For the *AG* subfamily, we found two *Gunnera* sequences orthologous to *Vitis* genes positioned on collinear gamma regions on chromosomes 10 and 12. For the *DEF* subfamily, we found that a previously cloned *GunneraTM6* gene is orthologous to *Vitis TM6* on chromosome 4, in a region collinear to chromosome 18 in which *Vitis euAP3* is located (Kramer et al. 2006). Three *Gunnera SQUA* genes are orthologous to *Vitis euAP1*, *euFUL*, and *AGL72* sequences located in gamma-derived regions on *Vitis* chromosomes 1, 14, and 17. The *Vitis SEP* genes are positioned next to these and for the *SEP4* (also *AGL3*) gene, orthology could be shown to a *Gunnera SEP4* sequence. For these first four subfamilies, Trochodendrales and Buxales sequences branch off before duplication or triplication, indicating that the gene duplications occurred in the stem lineage of core eudicots and derive from the gamma triplication. Two more phylogenies for the *SOC1* subfamily and the *AGL6* subfamily did not include Trochodendrales sequences, but only Buxales. In both, a single *Gunnera* sequence could be shown to be orthologous to a single *Vitis* sequence located in gamma-derived chromosomal segments. The combination of orthology relationships of *Gunnera* and *Vitis* genes in gamma-derived duplicate or triplicate regions thus establishes that these MADS-box gene duplications occurred in the stem lineage of core eudicots and derive from the gamma hexaploidization.

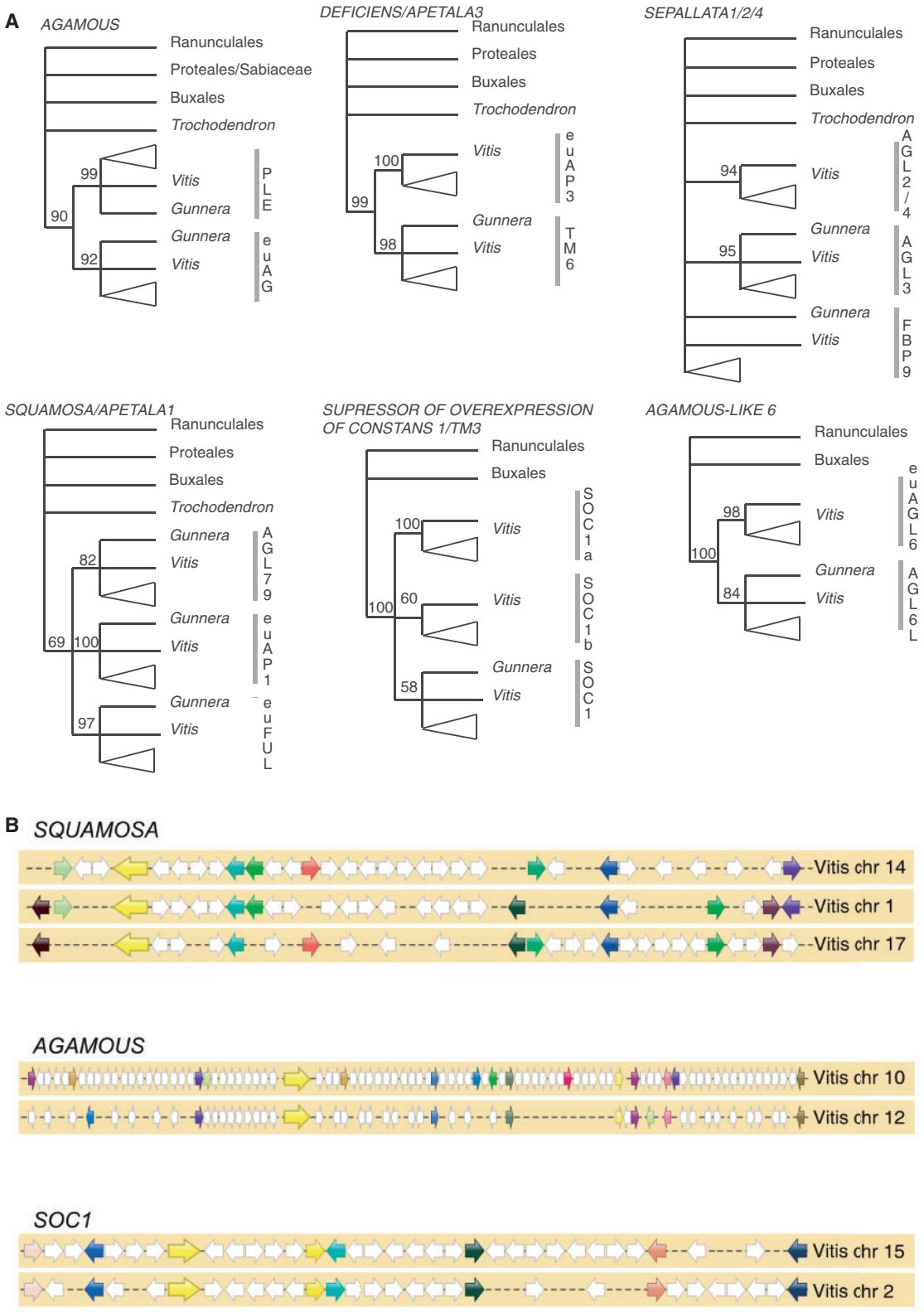
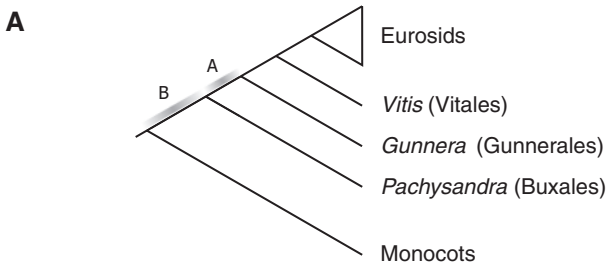


Fig. 2. (*A*) Gamma derived duplications and triplications in MADS-box gene subfamilies. Simplified phylogenetic trees summarizing results in supplementary figures S1–S6, Supplementary Material online, showing the supported topology of gamma orthologs between *Gunnera* and *Vitis*. Numbers are bootstrap values under the Maximum Likelihood criterion. Triangles represent other core eudicot lineages. (*B*) Gene order alignment for collinear regions in which *AGAMOUS*, *SQUAMOSA*, and *SOC1* are located. To show that MADS-box genes were indeed duplicated during the hexaploidization and not derived from independent small scale duplication events in the ancestor of the core eudicots, collinear regions in *Vitis* were analyzed. In these gene order alignments, genes retained in duplicate after the hexaploidization are shown in the same color, whereas white genes have no homologs in these stretches. Although relatively few genes have been retained after duplication, the conservation of both gene content and order of MADS-box genes (indicated by larger yellow arrows) is still evident, representative of a large-scale duplication event.

Phylogenomic Analysis of Pachysandra (Buxales) and Gunnera (Gunnerales) Transcriptomes Provides Evidence for the Gamma Triplication in the Stem Lineage of Core Eudicots

Using K_S, phylogenetic, and collinearity approaches (Van de Peer 2004), evidence for the gamma triplication has been found in most sequenced core eudicots (e.g., Jaillon et al. 2007; Ming et al. 2008; Cenci et al. 2010; Velasco et al. 2010; Xu et al. 2011; Sato et al. 2012). In the absence of a full genome, gene-collinearity methods like i-ADHoRe (Proost et al. 2012) and MCSCanX (Wang et al. 2012) cannot be used, as they require exact positional information of genes. However, after assembling reads into transcripts and annotating open reading frames, the resulting coding sequences can be clustered into gene families from which the paranome (collection of duplicated genes) can be studied further.

To verify our MADS-box gene-based phylogenetic placement of the gamma polyploidy event, we constructed orthogroups using orthoMCL from which we subsequently generated phylogenetic trees using PhyML (Guindon et al. 2010), which where automatically reconciled using NOTUNG (Chen et al. 2000). This resulted in an initial set of 14,397 reconciled trees. From this set, trees with a duplication node in the Eudicotyledoneae or core eudicots with bootstrap support of 80 or 50 were selected and compared with two hypothetical tree topologies as presented in figure 3. Scenario A corresponds to a situation where the gene duplicates are shared between the Gunnerales and Vitales, but not the Buxales, whereas in scenario B, the gene duplicates are also shared with the Buxales. To avoid including trees where a duplication is observed after reconciliation due to a slight difference between the gene and species tree, the duplication consistency score is calculated for all considered duplication nodes and nodes are only considered valid if a score >0.5 is found (Vilella et al. 2009). This measure avoids counting spurious duplication nodes that nevertheless often still have a high bootstrap value. As such, we counted 983 trees (93.26%) supporting a duplication in the MRCA of all core eudicots (scenario A) and only 71 trees (6.74%) supporting a duplication in the MRCA of core eudicots + Buxales (scenario B) with bootstrap support (BS) >80. We found similar results for BS support >50, with 1,718 (95.13%) trees supporting scenario A and 88 (4.87%) trees supporting scenario B. Using the 80 and 50 minimal bootstrap criterion, we also found 17 and 29 trees, respectively, that confirmed both scenarios. These have been excluded from further analyses. Furthermore, 75 (BS ≥ 80) and 123 (BS ≥ 50) trees supporting scenario A in fact contained *Vitis* genes known to be derived from gamma, based on collinearity (fig. 3*B*). Together, these results provide strong support for the gamma duplication to have occured after the Buxales branched off, but prior to the divergence of the Gunnerales.



B

	Bootstrap 80		Bootstrap 50	
	A	B	A	B
Supporting trees	983	71	1718	88
Percentage	93,26%	6,74%	95,13%	4,87%
Contains gamma derived duplicates (Vitis vinifera)	75		123	

Fig. 3. Hypothetical tree topologies and overview of gene trees consistent with either scenario. (*A*) Highly pruned phylogenetic tree indicating two timepoints where gamma could have occurred. The first scenario, indicated by A, is before the divergence of the Buxales and Gunnerales. Hence, the triplication is only present in *Gunnera* and the remaining core eudicots. The second possibility, indicated by B, places gamma however before the divergence of the Buxales. (*B*) Overview of how many trees support either scenario A or B with high and low bootstrap support (80 and 50, respectively). The vast majority of trees support scenario A, whereas only very limited support for scenario B was found. Furthermore, various trees supporting scenario A contained *Vitis vinifera* genes known to be derived from gamma based on collinearity.

K_S-Based Analysis Provides Further Support for the Gamma Triplication in the Stem Lineage of Core Eudicots

K_S-based age distributions were constructed and are presented in figure 4 for all duplicated gene pairs in *Pachysandra*, *Gunnera*, and *Vitis*. Synonymous substitutions are thought to be putatively neutral (Kimura 1977), and therefore to accumulate at an approximately constant rate. Consequently, the number of synonymous substitutions per synonymous site (K_S) between paralogs is often used as a proxy for time since duplication. WGDs result in sudden bursts of new gene duplicates concentrated in time, recognizable as spikes superimposed on an exponential decay distribution of small-scale duplications (SSDs) in K_S-based age distributions (Blanc and Wolfe 2004). To elucidate whether observed peaks in the age distributions correspond to real WGD events, we employed a mixture modeling approach (Schlueter et al. 2004; Cui et al. 2006) to identify the mixture of log-transformed SSD background exponentials and WGD Gaussian density components present in the empirical age distributions. It is known however, that the model selection criteria used to identify the optimal number of mixture components, such as BIC (Schwarz 1978), are prone to overfitting (see Materials and Methods and supplementary table S3, Supplementary Material online) (Naik et al. 2007). We therefore used the SiZer software (Chaudhuri and Marron 1999) to distinguish true features from noise in the age distributions (fig. 4). Results are summarized in table 1.

For *Pachysandra*, five components were fitted to the age distribution but SiZer analysis indicated none of these

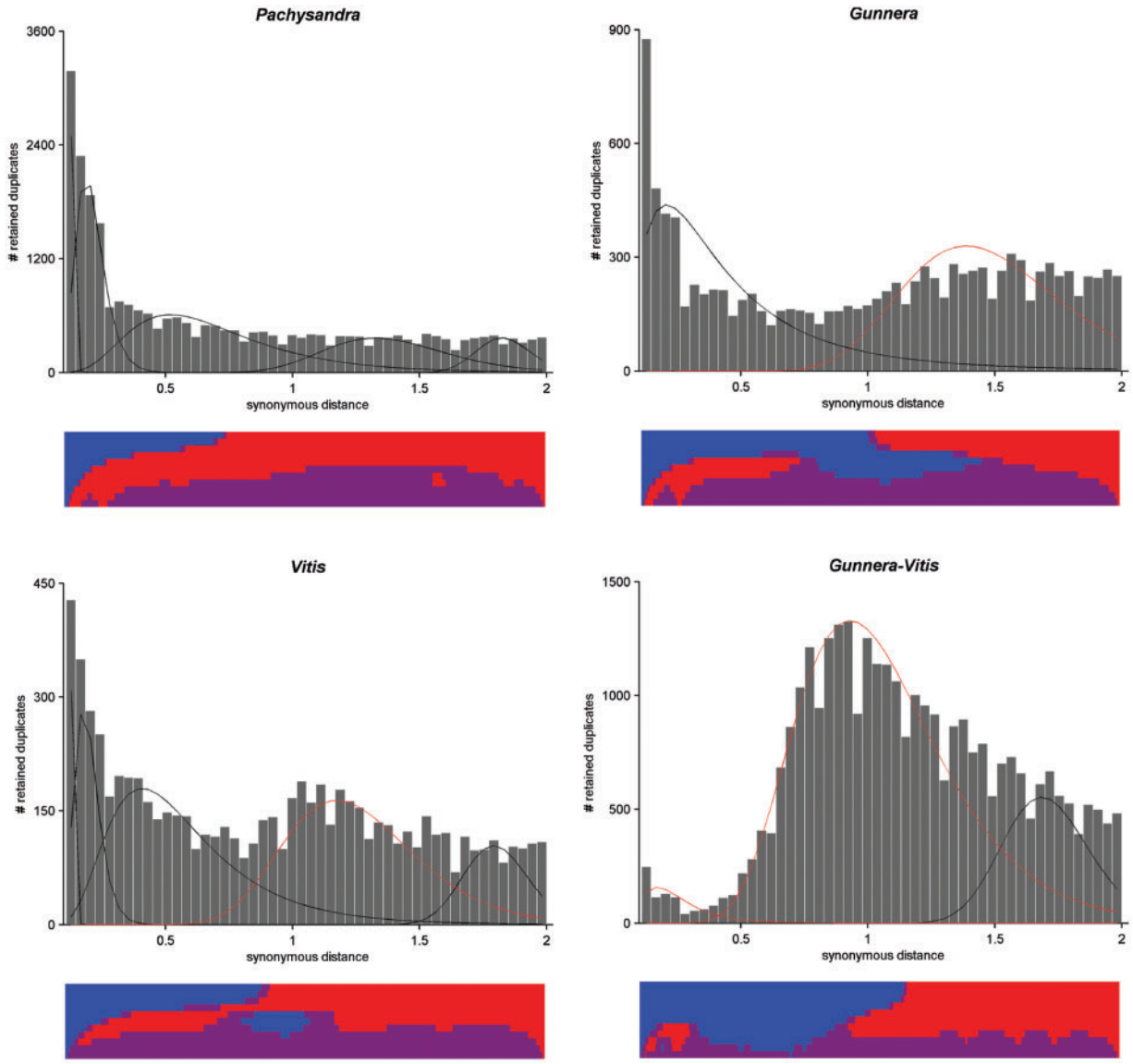


FIG. 4. K_S-based age distributions provide support for a shared gamma duplication event in the ancestor of core eudictos. Age distributions for the paranomes of *Pachysandra*, *Gunnera*, *Vitis*, and the *Gunnera–Vitis* orthologs are presented as indicated on top of the individual panels. SiZer results are presented underneath the corresponding age distributions. Components of the EMMIX mixture model corresponding to significant WGD features as identified by SiZer are plotted on the age distributions in red, whereas other components are plotted in black. A WGD event is identified in both *Gunnera* and *Vitis* around a K_S of 1.5 and 1.3, respectively, but not in *Pachysandra*. Evaluation of the *Gunnera–Vitis* ortholog plot indicates the speciation event took place around a K_S of 1.0, placing the WGD event in both species in time before the speciation event. The WGD peak in the *Gunnera* age distribution thus corresponds to the gamma duplication in *Vitis* age distribution.

correspond to a truly significant WGD feature. In contrast, four and five components were fitted to the age distributions of *Gunnera* and *Vitis*, respectively, of which SiZer analysis indicated one component in each species corresponds to a significant feature at a K_S of 1.3 and 1.5, respectively. These results indicate that a WGD occurred in the ancestor of both the Gunnerales and Vitales, but not in the ancestor of the Buxales. The peak in the *Vitis* distribution has already been shown before to correspond to the gamma duplication (Jiao et al. 2012). To investigate whether this was a shared WGD event between *Gunnera* and *Vitis*, we also constructed a K_S-based age distribution of the *Gunnera–Vitis* orthologs, where a peak represents the sudden creation of a set of orthologs concentrated in time, that is, a speciation event. Five components were fitted, of which SiZer analysis indicated two correspond to significant speciation features at a K_S of 0.2 and 1.0. The large peak of orthologs at a K_S of 1.0 represents the speciation event between *Vitis* and *Gunnera*, whereas the smaller peak at a K_S of 0.2 may represent a subset of orthologs created by the same speciation event that exhibit slower synonymous evolution,for example, because of gene conversion events with other paralogs. The fact that the ortholog peak is situated at a lower K_S value than both of the WGD peaks in *Gunnera* and *Vitis* indicates that the associated WGD event(s) predated the speciation event, and hence that the observed WGD peak in the *Gunnera* distribution corresponds to the observed gamma WGD peak in the *Vitis* distribution.

Table 1. Mixture Modeling and SiZer Analysis of the K_s-Based Age Distributions Presented in Figure 4.

Species	No. of Duplicates	No. of Components	BIC	Mixture Distribution Means (K_s)	Variance (K_s)	Proportion	Inferred Duplication
Pachysandra	28,239	5	60,874.62	0.124	0.0001	0.089	
				0.210	0.0035	0.260	
				0.702	0.1091	0.366	
				1.391	0.0619	0.200	
				1.833	0.0111	0.085	
Gunnera	11,970	4	20,198.25	0.484	0.1788	0.468	
				1.491	0.1091	0.527	Gamma
				1.520	0.0000	0.003	
				1.520	0.0000	0.003	
Vitis	7,370	5	14,592.62	0.123	0.0001	0.042	
				0.196	0.0026	0.119	
				0.575	0.0847	0.352	
				1.258	0.0748	0.376	Gamma
				1.805	0.0163	0.111	
Gunnera-Vitis	31,953	5	16,743.90	0.248	0.0173	0.029	Gamma
				1.059	0.1018	0.781	Gamma
				1.520	0.0000	0.003	
				1.520	0.0000	0.003	
				1.710	0.0297	0.185	

NOTE.—Properties of the components identified by EMMIX are indicated. Components corresponding to significant peaks in the age distribution as identified by SiZer are interpreted in light of their paleopolyploid history in the last column (see Results).

The *Gamma* Triplication Occurred Early After the Divergence of Buxales

Genome duplications often seem to be correlated with an increase in species richness (Otto and Whitton 2000; Vamosi and Dickinson 2006; Van de Peer et al. 2009). Similarly, here we placed the gamma event in the stem lineage of a large group of species, the core eudicots. This might suggest that polyploidy may be correlated with a rapid initial diversification, as has been suggested previously for both ancient (Bowers et al. 2003; de Bodt et al. 2005) and recent WGDs (Soltis et al. 2009). Whereas the correspondence in time between genome duplication and species diversification is more difficult to evaluate for older palaeopolyploidy events, because the absolute age of nodes is uncertain, it can be assessed more accurately for more recent events (Fawcett et al. 2009; Jiao et al. 2011). To evaluate in more detail, potential links between the gamma triplication and the diversification of extant core eudicot lineages in more detail, we estimated whether the gamma triplication occurred early or late in the stem lineage of core eudicots.

We used the six phylogenies of MADS-box genes to investigate this, following two approaches (table 2). In either, we constrained the eudicot diversification to 125 Mya, in accordance to the observation of tricolpate pollen in the fossil record (Magallón and Castillo 2009). For the first approach, we included several primary fossil calibration points in the phylogeny (supplementary table S2, Supplementary Material online). This resulted in the following time estimates: 122.36 ± 0.68 Mya [mean, standard deviation {SD}] for the divergence of the Buxales, 120.05 ± 2.61 Mya for the gamma triplication, and 113.43 ± 3.54 Mya for the divergence of the Gunnerales. This suggests that the divergence of the Gunnerales clade from the stem lineage occurred ~7 My after the palaeopolyploidy event. These first estimates depended mainly on the number and age of the included primary calibration points. To avoid such effects, in the second approach, the speciation of Gunnerales was fixed to 116.74 Mya (Magallón and Castillo 2009), leading to the following time estimates: 122.35 ± 0.93 Mya for the divergence of the Buxales, and 120.43 ± 1.50 Mya for the gamma triplication. This again suggests that the gamma triplication occurred early in the stem lineage of core eudicots and that the first extant species lineage diverged considerably later in time.

Discussion

Many questions relating to the role and significance of polyploidy in flowering plant diversity remain to be resolved (Otto 2007; Soltis et al. 2009; Van de Peer et al. 2009; Mayrose et al. 2011). Although Jiao et al. (2012) already provided strong evidence that gamma occurred before the divergence of asterids and rosids, and most likely after the earliest diversification of eudicots, our inclusion of the transcriptomes of *Pachysandra* (Buxales) and *Gunnera* (Gunnerales), occupying two crucial phylogenetic positions, enabled us to provide the most accurate phylogenetic placement of the gamma hexaploidy event to date. In addition, we demonstrate that retained duplicates of core eudicot MADS-box gene transcription factors derive from this event. The accurate placement in absolute time and relative to speciation events in the phylogeny allows evaluation of the impact of this ancient WGD event for species diversification and functional diversification of regulatory processes.

Table 2. Estimated Ages in Million Years of Crown Clades of Interest for Six MADS-Box Gene Lineages.

Crown Clade	*DEF*	*SQUA*	*AG*	*AGL6*	*SOC1*	*AGL2/3/4*	Mean (SD)
Approach 1							
MRCA *Buxales Trochodendrales*	121.57	123.2	122.62	124.31	120.25	123.68	122.61 (1.49)
MRCA *Buxales Gunnerales*	121.52	122.91	122.08	na	na	122.91	122.36 (0.68)
MRCA core eudicots (gamma triplication)	117.77	121.45	119.51	123.54	116.53	121.48	120.05 (2.61)
MRCA *Gunnerales Malpighiales* 1	110.76	109.59	107.14	119.06	112.87	117.16	113.43 (3.54)
MRCA *Gunnerales Malpighiales* 2	na	114.86	114.68	na	na	114.73	
MRCA *Gunnerales Malpighiales* 3	na	na	na	na	na	na	
Approach 2							
MRCA *Buxales Trochodendrales*	121.47	123.81	122.3	120.24	121.02	122.78	121.94 (1.29)
MRCA *Buxales Gunnerales*	121.44	123.64	122.08	na	na	122.25	122.35 (0.93)
MRCA core eudicots (gamma triplication)	119.41	122.75	120.92	119.71	118.57	121.24	120.43 (1.50)
MRCA *Gunnerales Malpighiales* 1	116.74	116.74	116.74	116.74	116.74	116.74	116.74
MRCA *Gunnerales Malpighiales* 2	na	116.74	116.74	na	na	116.74	
MRCA *Gunnerales Malpighiales* 3	na	116.74	na	na	na	na	

Note.—Age estimates using two different approaches are listed (see Materials and Methods for age constraints). na, not available.

We find that, rather than immediately before Gunnerales branch off, the gamma triplication probably occurred early in the core eudicot stem lineage. Whereas a correlation between species diversity and polyploid origin has been observed (Otto and Whitton 2000; Vamosi and Dickinson 2006; Magallón and Castillo 2009; Soltis et al. 2009), our data, in agreement with recent observations for more recent polyploids (Mayrose et al. 2011), suggest that the ancestral polyploid did not rapidly generate extant species lineages. It is possible that the accuracy of these dating estimates is affected by MADS-box gene diversification after duplication or by methodological limitations. Also, the precision of these estimates could be further improved by including more gene families in our analysis. Nevertheless, these estimates for independent gene trees suggest that if the polyploidy event contributed to core eudicot diversification, this was only realized after a few million years. This could also be true for the paleoploidy event epsilon in the flowering plant stem lineage (Jiao et al. 2011). Based on similar observations in important angiosperm crowngroups (e.g., Brassicaceae, Poaceae, Asteraceae, Fabaceae, and Solanaceae), Schranz et al. (2012) recently proposed a WGD radiation lag-time (WGD-RTL) model. In this model, WGDs are still important for the evolution of novel key traits, but the ultimate succes of a phylogenetic group lies mainly in subsequent (millions of years later) evolutionary events including migration, changing environmental conditions, and/or differential extinction rates (Schranz et al. 2012). Related to this, also in ray-finned fishes, a significant delay in species diversification following WGDs was observed (Santini et al. 2009; Van de Peer et al. 2009).

During this lag-time period, a higher resilience against extinction could have contributed to polyploid survival (Crow and Wagner 2006). Similar to WGDs that appear to have occurred close to the K-T extinction event 60–70 Mya (Fawcett et al. 2009), the gamma triplication occurred close in time to the proposed early Aptian extinction event 120 Mya (Archangelsky 2001). This land extinction event is associated with the better known Oceanic anoxic event 1a (Schlanger and Jenkins 1976; Erba et al. 2010). The latter was initiated by volcanic activity and associated with fluctuating carbon dioxide levels, high temperatures, and arid conditions that resulted in shifts in vegetation (Larson and Erba 1999, Keller et al. 2011).

In accordance with the WGD-RLT model, after what appears to be an extended period of survival, the hexaploid stem lineage of core eudicots suddenly generated many lineages with several extant representatives (Moore et al. 2010). Possibly, the Aptian extinction would have cleared ecological niches available for gamma descendants to occupy, unfolding their intrinsic potential acquired through the gamma event (Mayrose et al. 2011).

Genome duplication has been proposed as a driver of morphological complexity (Freeling and Thomas 2006). Like homeobox genes in animals, MADS-box genes have become principal targets of studies in plants that focus on the understanding of developmental evolution (Doebley and Lukens 1998; Cañestro et al. 2007; Theissen and Melzer 2007; Lee and Irish 2011). We have shown here that MADS-box gene lineages within subfamilies expanded following the gamma triplication and were retained in subsequent core eudicot evolution. This provides a first angle to study the functional implications of ancient WGDs in plants. The gamma triplication, through functional diversification of regulatory genes, could have been instrumental to the origin and diversification of core eudicots. For instance, the duplication in the *DEF* subfamily may be related to the origin of a clearer separation of sepal and petal identity in core eudicots (e.g., Hileman and Irish 2009), the duplication in the *AGL6* subfamily may have resulted in expression in vegetative tissue (Viaene et al. 2010) and changes in protein interactions could be related to functional diversification in the *SQUA* subfamily (Liu et al. 2010). To better infer ancestral functions of MADS-box genes and how they have diversified, more functional data in phylogenetically informative species are required (Hileman and Irish 2009). The development of transient functional methods applicable to many species, such as virus-induced gene silencing,

will contribute to make more confident inferences of the functional impact of gene and genome duplications (Hileman et al. 2005; Drea et al. 2007; Kramer et al. 2007; Becker and Lange 2010). It will be interesting to study how fast these gene lineage specific functions were established during evolution, as this will contribute to our understanding of the genetic potential of the gamma triplication and how it was realized.

An alternative approach to understand the impact of the gamma triplication is to start from the derived traits of core eudicots as these characters are likely candidates to have played a major role in the establishment of this group of species. Although few derived traits characterize core eudicots in a strict sense, several characters appear to have originated early in core eudicot evolution. The best known derived character shared by Gunnerales and other core eudicots is the presence of gallo- and ellagitannins (Bate-Smith 1962, 1968; Doyle 1988; Okuda et al. 2000). These secondary metabolites function both as anti-oxidants and microbicides. A recent study identified shikimate dehydrogenase as the enzyme directly responsible for the synthesis of gallic acid in plants (Muir et al. 2011). Interestingly, we found that this gene lineage indeed underwent a triplication at the origin of core eudicots of which only one lineage is retained in *A. thaliana* (supplementary fig. S7). Future research should establish whether the gamma triplication was indeed instrumental for the origin of this trait.

Supplementary Material

Supplementary methods, figures S1–S7, and tables S1–S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We would like to thank Anja Vandeperre and Nathalie Geerts for technical support, Leentje Van Lommel (Gene Expression Group, KULeuven) for performing Agilent Bioanalyzer measurements, and Elke Bellefroid from the National Botanical Garden of Belgium for the collection of material. This work was supported by Fonds voor Wetenschappelijk Onderzoek Vlaanderen (FWO) project (G.0607.11 N), by Ghent University (Multidisciplinary Research Partnership "Bioinformatics: from nucleotides to networks") and Research Fund KU Leuven project 12/053 awarded to K.G.; the Interuniversity Attraction Poles Programme (IUAP P6/25), initiated by the Belgian State, Science Policy Office (BioMaGNet); and the European Union Framework Program "Food Safety and Quality" (FOOD-CT-2006-016214). D.V., K.V., and S.M. acknowledge FWO scholarships; H.C. and T.V. were supported by the KULeuven; and P.R. acknowledges a scholarship from the agentschap voor Innovatie door Wetenschap en Technologie (IWT).

References

Adams KL, Wendel JF. 2005. Polyploidy and genome evolution in plants. *Curr Opin Plant Biol.* 8:135–141.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389–3402.

Angiosperm Phylogeny Group. 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *J Linn Soc Lond Bot.* 161:105–121.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408(6814):796–815.

Archangelsky S. 2001. The Ticó Flora (Patagonia) and the Aptian Extinction Event. *Acta Palaeobot.* 41:115–122.

Baker M. 2012. De novo genome assembly: what every biologist should know. *Nat Methods.* 9:333–337.

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.

Bate-Smith EC. 1962. The phenolic constituents of plants and their taxonomic significance. I. Dicotyledons. *J Linn Soc Lond Bot.* 58: 95–173.

Bate-Smith EC. 1968. The phenolic constituents of plants and their taxonomic significance. *J Linn Soc Lond Bot.* 60:325–356.

Becker A, Lange M. 2010. VIGS—genomics goes functional. *Trends Plant Sci.* 15:1–4.

Birchler JA, Veitia RA. 2010. The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. *New Phytol.* 186:54–62.

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

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

Brenner GJ. 1996. Evidence of the earliest stage of angiosperm pollen evolution: a paleoequatorial section from Israel. In: Taylor DW, Hickey LJ, editors. Flowering plant origin, evolution, and phylogeny. New York: Chapman and Hall. p. 91–115.

Cañestro C, Yokoi H, Postlethwait JH. 2007. Evolutionary developmental biology and genomics. *Nat Rev Genet.* 8:932–942.

Cenci A, Combes MC, Lasthermes P. 2010. Comparative sequence analyses indicate that *Coffea* (Asterids) and *Vitis* (Rosids) derive from the same paleo-hexaploid ancestral genome. *Mol Genet Genomics.* 283:493–501.

Chaudhuri P, Marron J. 1999. SiZer for exploration of structures in curves. *J Am Stat Assoc.* 94:807–823.

Chen K, Durand D, Farach-Colton M. 2000. NOTUNG: a program for dating gene duplications and optimizing gene family trees. *J Comp Biol.* 7:429–447.

Childs KL, Hamilton JP, Zhu W, Ly E, Cheung F, Wu H, Rabinowicz PD, et al. (10 co-authors). 2007. The TIGR Plant Transcript Assemblies Database. *Nucleic Acids Res.* 35(Database issue):D846–D851.

Crow KD, Wagner GP. 2006. What is the role of genome duplication in the evolution of complexity and diversity? *Mol. Biol. Evol.* 23: 887–892.

Cui L, Wall PK, Leebens-Mack JH, et al. (13 co-authors). 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Res.* 16:738–749.

De Bodt S, Maere S, Van de Peer Y. 2005. Genome duplication and the origin of angiosperms. *Trends Ecol Evol.* 20:591–597.

De Craene LPR. 2007. Are petals sterile stamens or bracts? The origin and evolution of petals in the core eudicots. *Ann Bot.* 100: 621–630.

Díaz-Riquelme J, Lijavetzky D, Martínez-Zapater JM, Carmona MJ. 2009. Genome-wide analysis of MIKCC-type MADS-box genes in grapevine. *Plant Physiol.* 149:354–369.

Doebley J, Lukens L. 1998. Transcriptional regulators and the evolution of plant form. *Plant Cell.* 10:1075–1082.

Doyle M. 1988. A comparative phytochemical profile of the Gunneraceae. *NZ J Bot.* 26:493–496.

Drea S, Hileman LC, de Martino G, Irish VF. 2007. Functional analyses of genetic pathways controlling petal specification in poppy. *Development* 134:4157–4166.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32: 1792–1797.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30: 1575–1584.

Erba E, Bottini C, Weissert HJ, Keller CE. 2010. Calcareous nannoplankton response to surface-water acidification around Oceanic Anoxic Event 1a. *Science* 329:428–432.

Fawcett JA, Maere S, Van de Peer Y. 2009. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proc Natl Acad Sci U S A.* 106:5737–5742.

Fawcett JA, Van de Peer Y. 2010. Angiosperm polyploids and their road to evolutionary success. *Trends Evol Biol.* 2:e3.

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.

Geuten K, Irish V. 2010. Hidden variability of floral homeotic B genes in Solanaceae provides a molecular basis for the evolution of novel functions. *Plant Cell.* 22:2562–2578.

Goff SA, Ricke D, Briggs S. 2002. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science* 296:92–100.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol.* 11: 725–736.

Gouzy J, Carrere S, Schiex T. 2009. FrameDP: sensitive peptide detection on noisy matured sequences. *Bioinformatics* 25:670–671.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol.* 5:307–321.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52: 696–704.

Hakes L, Pinney JW, Lovell SC, Oliver SG, Robertson DL. 2007. All duplicates are not equal: the difference between small-scale and genome duplication. *Genome Biol.* 8:R209; 201–R209.213.

Hileman LC, Drea S, Martino G, Litt A, Irish VF. 2005. Virus-induced gene silencing is an effective tool for assaying gene function in the basal eudicot species Papaver somniferum (opium poppy). *Plant J.* 44: 334–341.

Hileman LC, Irish VF. 2009. More is better: the use of developmental genetic data to reconstruct perianth evolution. *Am J Bot.* 96: 83–95.

Hughes NF, McDougall AB. 1990. Barremian-Aptian angiospermid pollen records from southern England. *Review of Palaeobotany and Palynology* 65:145–151.

Hughes NF, McDougall AB, Chapman JL. 1991. Exceptional new record of Cretaceous Hauterivian angiospermid pollen from southern England. *Journal of Micropalaeontology* 10:75–82.

Irish VF. 2006. Duplication, diversification, and comparative genetics of angiosperm MADS-box genes. *Adv Bot Res.* 44:129–161.

Irish VF, Litt A. 2005. Flower development and evolution: gene duplication, diversification and redeployment. *Curr Opin Genet Dev.* 15: 454–460.

Jaillon O, Aury J-M, Noel B, et al. (57 co-authors). 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449:463–467.

Jiao Y, Leebens-Mack J, Ayyampalayam S, et al. (25 co-authors). 2012. A genome triplication associated with early diversification of the core eudicots. *Genome Biol.* 13:R3.

Jiao Y, Wickett NJ, Ayyampalayam S, et al. (17 co-authors). 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473: 97–100.

Kaufmann K, Melzer R, Theissen G. 2005. MIKC-type MADS-domain proteins: structural modularity, protein interactions and network evolution in land plants. *Gene* 347:183–198.

Keller CE, Hochuli PA, Weissert H, Bernasconi SM, Giorgioni M, Garcia TI. 2011. A volcanically induced climate warming and floral change preceded the onset of OAE1a (Early Cretaceous). *Palaeogeogr Palaeoclimatol Palaeoecol.* 305:43–49.

Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* 428:617–624.

Kimura M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* 267(5608):275–276.

Kramer EM, Dorit RL, Irish VF. 1998. Molecular evolution of genes controlling petal and stamen development: duplication and divergence within the APETALA3 and PISTILLATA MADS-box gene lineages. *Genetics* 149:765–783.

Kramer EM, Holappa L, Gould B, Jaramillo MA, Setnikov D, Santiago PM. 2007. Elaboration of B gene function to include the identity of novel floral organs in the lower eudicot Aquilegia. *Plant Cell.* 19: 750–766.

Kramer EM, Su H-J, Wu C-C, Hu J-M. 2006. A simplified explanation for the frameshift mutation that created a novel C-terminal motif in the APETALA3 gene lineage. *BMC Evol Biol.* 6:30.

Larson R, Erba E. 1999. Onset of the mid-Cretaceous greenhouse in the Barremian-Aptian: igneous events and the biological, sedimentary, and geochemical responses. *Paleoceanography* 14: 663–678.

Lee H-L, Irish VF. 2011. Gene duplication and loss in a MADS-box gene transcription factor circuit. *Mol Biol Evol.* 28:3367–3380.

Levin DA. 1983. Polyploidy and novelty in flowering plants. *Am Nat.* 122: 1–25.

Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genoms. *Genome Res.* 13:2178–2189.

Li R, Zhu H, Wang J. 14. 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* 20: 265–272.

Litt A, Irish VF. 2003. Duplication and diversification in the APETALA1/FRUITFULL floral homeotic gene lineage: implications for the evolution of floral development. *Genetics* 165:821–833.

Liu B, Wendel JF. 2003. Epigenetic phenomena and the evolution of plant allopolyploids. *Mol Phylogen Evol.* 29:365–379.

Liu C, Zhang J, Zhang N, Shan H, Su K, Zhang J, Meng Z, Kong H, Chen Z. 2010. Interactions among proteins of floral MADS-box genes in basal eudicots: implications for evolution of the regulatory network for flower development. *Mol Biol Evol.* 27:1598–1611.

Maclean CJ, Greig D. 2011. Reciprocal gene loss following experimental whole-genome duplication causes reproductive isolation in yeast. *Evolution* 65:932–945.

McLachlan G, Peel D, Basford K, Adams P. 1999. The EMMIX software for the fitting of mixtures of normal and t-components. *J Stat Softw.* 4:2.

Maddison DR, Maddison WP. 2000. MacClade 4: analysis of phylogeny and character evolution. Version 4. Sunderland (MA): Sinauer Associates.

Maere S, De Bodt S, Raes J, Casneuf T, Van 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.

Magallón SS, Castillo AA. 2009. Angiosperm diversification through time. *Am J Bot.* 96:349–365.

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* 333:1257–1257.

Ming R, Hou S, Feng Y, et al. (84 co-authors). 2008. The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). *Nature* 452:991–996.

Moore MJ, Hassan N, Gitzendanner MA, et al. (17 co-authors). 2011. Phylogenetic analysis of the plastid inverted repeat for 244 species: insights into deeper-level angiosperm relationships from a long, slowly evolving sequence region. *Int J Plant Sci.* 172(4):541–558.

Moore MJ, Soltis PS, Bell CD, Burleigh JG, Soltis DE. 2010. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proc Natl Acad Sci U S A.* 107:4623–4628.

Muir RM, Ibáñez AM, Uratsu SL, et al. (11 co-authors). 2011. Mechanism of gallic acid biosynthesis in bacteria (Escherichia coli) and walnut (Juglans regia). *Plant Mol Biol.* 75:555–565.

Naik PA, Shi P, Tsai CL. 2007. Extending the akaike information criterion to mixture regression models. *J Am Stat Assoc.* 102:244–254.

Ohno S. 1970. Evolution by gene duplication. Berlin/Heidelberg: Springer-Verlag.

Okuda T, Yoshida T, Hatano T. 2000. Correlation of oxidative transformations of hydrolyzable tannins and plant evolution. *Phytochemistry* 55:513–529.

Otto SP. 2007. The evolutionary consequences of polyploidy. *Cell.* 131: 452–462.

Otto SP, Whitton J. 2000. Polyploid incidence and evolution. *Annu Rev Genet.* 34:401–437.

Posada D, Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14:817–818.

Proost S, Fostier J, De Witte D, Dhoedt B, Demeester P, Van de Peer Y, Vandepoele K. 2012. i-ADHoRe 3.0–fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Res.* 40:e11.

Proost S, Van Bel M, Sterck L, Billiau K, Van Parys T, Van de Peer Y, Vandepoele K. 2009. PLAZA: a comparative genomics resource to study gene and genome evolution in plants. *Plant Cell.* 21:3718–3731.

Ramsey J. 2011. Polyploidy and ecological adaptation in wild yarrow. *Proc Natl Acad Sci U S A.* 108:7096–7101.

Ramsey J, Schemske DW. 1998. Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annu. Rev. Ecol. Syst.* 29(1):467–501.

Rothwell GW, Scheckler SE. 1988. Biology of ancestral gymnosperms. In: Beck CB, editor. Origin and evolution of gymnosperms. New York: Columbia University Press. p. 85–134.

Sanderson MJ. 2002. R8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19:301–302.

Santini F, Harmon LJ, Carnevale G, Alfaro ME. 2009. Did genome duplication drive the origin of teleosts? A comparative study of diversification in ray-finned fishes. *BMC Evol Biol.* 9:194.

Sato S, Tabata S, Hirakawa H, et al. (301 co-authors). 2012. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485:635–641.

Schlanger SO, Jenkyns HC. 1976. Cretaceous oceanic anoxic events: causes and consequences. Geologie en mijnbouw. 55:179–184.

Schlueter JA, Dixon P, Granger C, Grant D, Clark L, Doyle JJ, Shoemaker RC. 2004. Mining EST databases to resolve evolutionary events in major crop species. *Genome* 47:868–876.

Schranz ME, Mohammadin S, Edger PP. 2012. Ancient whole genome duplications, novelty and diversification: the WGD radiation lag-time model. *Curr Opin Plant Biol.* 15:147–153.

Schwarz G. 1978. Estimating the dimension of a model. *Ann Stat.* 6: 461–464.

Shan H, Zhang N, Liu C, Xu G, Zhang J, Chen Z, Kong H. 2007. Patterns of gene duplication and functional diversification during the evolution of the AP1/SQUA subfamily of plant MADS-box genes. *Mol Phylogenet Evol.* 44:26–41.

Simillion C, Vandepoele K, Van de Peer Y 5. 2002. The hidden duplication past of Arabidopsis thaliana. *Proc Natl Acad Sci U S A.* 99: 13627–13632.

Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, Sankoff D, dePamphilis CW, Kerr Wall P, Soltis PS. 2009. Polyploidy and angiosperm diversification. *Am J Bot.* 96:336–348.

Soltis DE, Senters AE, Zanis MJ, Kim S, Thompson JD, Soltis PS, De Craene LPR, Endress PK, Farris JS. 2003. Gunnerales are sister to other core eudicots: implications for the evolution of pentamery. *Am J Bot.* 90:461–470.

Soltis DE, Smith SA, Cellinese N, et al. (28 co-authors). 2011. Angiosperm phylogeny: 17 genes, 640 taxa. *Am J Bot.* 98:704–730.

Soltis PS, Soltis DE. 2000. The role of genetic and genomic attributes in the success. of polyploids. *Proc Natl Acad Sci U S A.* 97: 7051–7057.

Soltis DE, Soltis PS, Endress PK, Chase MW. 2005. Phylogeny and evolution of angiosperms. Sunderland, MA, USA: Sinauer Associates.

Stevens PF. 2001. Angiosperm Phylogeny Website. Version 9, June 2008 [and more or less continuously updated since].

Taylor JS, Braasch I, Van de Peer Y. 5. 2003. Genome duplication, a trait shared by 22,000 species of ray-finned fish. *Genome Res.* 13:382–390.

Theissen G, Melzer R. 2007. Molecular mechanisms underlying origin and diversification of the angiosperm flower. *Ann Bot.* 100: 603–619.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties, and weight matrix choice. *Nucleic Acids Res.* 22(22):4673–4680.

Tuskan GA, Difazio S, Jansson S, et al. (90 co-authors). 2006. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science* 313:1596–1604.

Vamosi JC, Dickinson TA. 2006. Polyploidy and diversification: A phylogenetic investigation in Rosaceae. *Int J Plant Sci.* 167:349–358.

Van Bel M, Proost S, Wischnitzki E, Movahedi S, Scheerlinck C, Van de Peer Y, Vandepoele K. 2012. Dissecting plant genomes with the PLAZA comparative genomics platform. *Plant Physiol.* 158:590–600.

Van de Peer Y, Maere S, Meyer A. 2009. The evolutionary significance of ancient genome duplications. *Nat Rev Genet.* 10:725–732.

Van de Peer Y. 2004. Computational approaches to unveiling ancient genome duplications. *Nat Rev Genet.* 5:752–763.

Velasco R, Zharkikh A, Affourtit J, et al. (86 co-authors). 2010. The genome of the domesticated apple (Malus x domestica Borkh.). *Nat Genet.* 42:833–839.

Veron AS, Kaufmann K, Bornberg-Bauer E. 2007. Evidence of interaction network evolution by whole-genome duplications: a case study in MADS-box proteins. *Mol Biol Evol.* 24:670–678.

Viaene T, Vekemans D, Becker A, Melzer S, Geuten K. 2010. Expression divergence of the AGL6 MADS domain transcription factor lineage after a core eudicot duplication suggests functional diversification. *BMC Plant Biol.* 10:148.

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* 19:D846–51.

Wang Y, Tang H, Debarry JD, et al. (12 co-authors). 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40:e49.

Winter KU, Becker A, Münster T, Kim JT, Saedler H, Theissen G. 1999. MADS-box genes reveal that gnetophytes are more closely related to conifers than to flowering plants. *Proc Natl Acad Sci U S A.* 96: 7342–7347.

Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387:708–713.

Wood TE, Takebayashi N, Rieseberg LH 6. 2009. The frequency of polyploid speciation in vascular plants. *Proc Natl Acad Sci U S A.* 106: 13875–13879.

Wu H, Su H. 2007. The identification of A-, B-, C-, and E-class MADS-box genes and implications for perianth evolution in the basal eudicot Trochodendron aralioides. *Int J Plant Sci.* 168:775–799.

Xu X, Pan S, Cheng S, et al. (98 co-authors). 2011. Genome sequence and analysis of the tuber crop potato. *Nature* 475:189–195.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.

Zahn LM, Leebens-Mack JH, Arrington JM, Hu Y, Landherr LL, dePamphilis CW, Becker A, Theissen G, Ma H. 2006. Conservation and divergence in the AGAMOUS subfamily of MADS-box genes: evidence of independent sub- and neofunctionalization events. *Evol Development.* 8:30–45.

Zahn LM, Kong H, Leebens-Mack JH, Kim S, Soltis PS, Landherr LL, Soltis DE, dePamphilis CW, Ma H. 2005. The evolution of the SEPALLATA subfamily of MADS-box genes: a preangiosperm origin with multiple duplications throughout angiosperm history. *Genetics* 169: 2209–2223.