RESEARCH ARTICLE

Paleopolyploidy in the Brassicales: Analyses of the *Cleome* Transcriptome Elucidate the History of Genome Duplications in *Arabidopsis* and Other Brassicales

Michael S. Barker,†[1] Heiko Vogel,‡ and M. Eric Schranz§[1]*

*Department of Botany and The Biodiversity Research Centre, University of British Columbia, Vancouver, Canada; †Department of Biology, Indiana University, Bloomington; ‡Max Planck Institute for Chemical Ecology, Jena, Germany; and §Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Amsterdam, The Netherlands

The analysis of the *Arabidopsis* genome revealed evidence of three ancient polyploidy events in the evolution of the Brassicaceae, but the exact phylogenetic placement of these events is still not resolved. The most recent event is called the *At*-α (alpha) or 3R, the intermediate event is referred to as the *At*-β (beta) or 2R, and the oldest is the *At*-γ (gamma) or 1R. It has recently been established that *At*-γ is shared with other Rosids, including papaya (*Carica*), poplar (*Populus*), and grape (*Vitis*), whereas data to date suggest that *At*-α is Brassicaceae specific. To address more precisely when the *At*-α and *At*-β events occurred and which plant lineages share these paleopolyploidizations, we sequenced and analyzed over 4,700 normalized expressed sequence tag sequences from the Cleomaceae, the sister family to the Brassicaceae. Analysis of these *Cleome* data with homologous sequences from other Rosid genomes (*Arabidopsis*, *Carica*, *Gossypium*, *Populus*, and *Vitis*) yielded three major findings: 1) confirmation of a *Cleome*-specific paleopolyploidization (*Cs*-α) that is independent of the Brassicaceae *At*-α paleopolyploidization; 2) *Cleome* and *Arabidopsis* share the *At*-β duplication, which is lacking from papaya within the Brassicales; and 3) rates of molecular evolution are faster for the herbaceous annual taxa *Arabidopsis* and *Cleome* than the other predominantly woody perennial Rosid lineages. These findings contribute to our understanding of the dynamics of genome duplication and evolution within one of the most comprehensively surveyed clades of plants, the Rosids, and clarify the complex history of the *At*-α, *At*-β, and *At*-γ duplications of *Arabidopsis*.

Introduction

Polyploidy, or whole-genome duplication (WGD), is widely recognized as an important mechanism of plant speciation and evolution (Doyle et al. 2008; Soltis et al. 2009). Recently derived polyploids, or neopolyploids, are easily identified compared with diploid progenitors because of increased genome size, higher chromosome numbers, and redundant gene content. Nearly 30% of extant flowering plants are neopolyploids (Wood et al. 2009), and recent studies have identified polyploidy throughout their evolutionary history (Soltis et al. 2009). However, these ancient polyploid events, or paleopolyploidizations, are more difficult to detect because of the rather enigmatic "diploidization" process (Freeling and Thomas 2006; Doyle et al. 2008). Diploidization refers to the suite of molecular mechanisms that ultimately leads to the return of many genes to single copy (gene fractionation), disomic chromosomal inheritance, and often to smaller genome size and reduced chromosome numbers. Several studies have shown that diploidization mechanisms, such as chromosomal translocations, deletion of repetitive sequences, and gene silencing, can begin immediately after polyploidization (e.g., Gaeta et al. 2007). Although we are aware of some mechanisms involved in diploidization, the rate of these changes and the exact phylogenetic placement of past polyploidizations are still poorly understood.

Following WGD, many genes return to single copy by fractionation, but some duplicate gene pairs are preferentially maintained. These are a special class of paralogous genes, variously referred to as homeologs, syntelogs (Freeling 2009), ohnologs (Wolfe 2000), or paleologs (Chapman et al. 2006). Analyses of paleologs have been central to the identification and relative dating of plant paleopolyploidizations. Paleologs may be used to identify collinearity and infer syntenic genomic regions in complete genome or bacterial artificial chromosome (BAC) sequences (Bowers et al. 2003; Jaillon et al. 2007; Ming et al. 2008). If the genome duplication was a tetraploidization, then there should be paired blocks of synteny, whereas an ancient hexaploidization (Schranz and Mitchell-Olds 2006; Town et al. 2006; Ming et al. 2008) yields trios of sytenic blocks.

When complete genomic sequences are not available, age distributions of gene duplications from transcriptome data, such as expressed sequence tags (ESTs), may be used to infer paleopolyploidy (Blanc and Wolfe 2004; Maere et al. 2005). WGD yields a common signal of duplication across multiple gene families because all genes are simultaneously duplicated. Hence, paleopolyploidizations appear as a significant enrichment in the age distribution of gene duplications. A variety of statistical methods have been applied to identify significant features in these age distributions, such as Kolmogorov–Smirnov (K-S) tests (Cui et al. 2006), SiZer (Barker et al. 2008), and mixture models (Schlueter et al. 2004), as well as place duplications in a relative phylogenetic framework (Barker et al. 2008). The relative divergence of the duplicates can then be used to infer the age of the duplication event.

To date, all fully sequenced flowering plant genomes, including monocots (*Oryza* and *Sorghum*) and eudicots (*Arabidopsis*, *Carica*, *Populus*, and *Vitis*), contain evidence

[1] These authors contributed equally to this work.

Key words: polyploidy, Brassicales, *Cleome*, *Arabidopsis*, transcriptome.

E-mail: m.e.schranz@uva.nl.

Genome. Biol. Evol. Vol. 2009:391–399.
doi:10.1093/gbe/evp040
Advance Access publication October 5, 2009

of at least one paleopolyploidization. Additionally, analyses of partial genome sequences (e.g., BACs) and ESTs have established polyploidy in the ancestry of many other families including the Solanaceae (Blanc and Wolfe 2004), Fabaceae (Blanc and Wolfe 2004; Schlueter et al. 2004), Compositae (Barker et al. 2008), and Cleomaceae (Schranz and Mitchell-Olds 2006). A key result of the sequencing of both *Carica* and *Vitis* is that they have undergone only one paleohexaploidy event, as evidenced by trios of syntenic regions across their genomes (Jaillon et al. 2007; Ming et al. 2008).

Genomic analyses have additionally revealed the signal of multiple paleopolyploid events of various ages within one genome. For example, the analysis of the *Arabidopsis* genome revealed evidence of at least three ancient polyploidy events. The most recent event with the largest set of maintained paleologs is called the *At*-α (alpha) or 3R, the intermediate event is referred to as the *At*-β (beta) or 2R, and the oldest, and with the fewest maintained paleologs, is the *At*-γ (gamma) or 1R. Two key questions arising from the analysis of the *Arabidopsis* genome are as follows: When precisely did these three events occur and which plant lineages share these paleopolypoidizations or potentially have their own independent genome duplications?

Initial analyses of paleopolyploidy in *Arabidopsis* relied on molecular clock estimates of the rate of synonymous substitutions (K_s) to date the age of the various events. Based on these calculations, *At*-α was thought to have occurred 14.5–86 Ma, *At*-β some 170–235 Ma, and the *At*-γ event nearly 300 Ma (Bowers et al. 2003). Hence, it was hypothesized that *At*-α might be as old as the order Brassicales, *At*-β might have occurred near the radiation of the eudicots, and *At*-γ might have occurred near the origin of angiosperms themselves (De Bodt et al. 2005). However, a recent study by Smith and Donoghue (2008) demonstrated that herbaceous species have a faster rate of molecular evolution than woody species. This finding has profound implications for the dating of paleopolyploidy events.

The initial dating of the *Arabidopsis* α, β, and γ events likely overestimated the age of these duplications because *Arabidopsis* is a herbaceous annual. Indeed, recent genomic analyses have borne this out. For example, *At*-γ is now recognized to be the same duplication as the paleohexapolyploidization detected in both *Carica* and *Vitis* (Lyons et al. 2008; Tang, Wang et al. 2008). Hence, this event is likely shared by all Rosids, and potentially all eudicots, but is likely not as old as the origin of the angiosperms. Secondly, the *Carica* genome did not contain evidence for having undergone *At*-β even though both *Arabidopsis* and *Carica* are members of the same order, the Brassicales. Hence, the *At*-β duplication is much younger than initially thought, but its exact phylogenetic position within the Brassicales is still not known. Finally, BAC sequencing of the herbaceous annual *Cleome spinosa*, a member of the family Cleomaceae, did not find evidence of the *At*-α event (Schranz and Mitchell-Olds 2006). This is significant because the Cleomaceae is sister to the Brassicaceae, and it restricts the *At*-α paleopolyploidization at or near the origin of the family Brassicaceae (the basal genus *Aethionema* may not share *At*-α with the rest of the family). Also of significance, the limited BAC sequencing of *Cleome* strongly suggested that it had undergone a more recent independent paleohexapolyploidization (Schranz and Mitchell-Olds 2006).

Further analysis of the *Cleome* genome has great potential to elucidate the role of paleopolyploidy in the evolution of the Brassicales. To facilitate *Cleome* genomics, we sequenced over 4,700 ESTs from the 5′-end of a normalized cDNA library. The analysis of these *Cleome* sequences and homologous sequences from other Rosid genomes (*Arabidopsis*, *Carica*, *Gossypium*, *Populus*, and *Vitis*) are used to address the following three issues: 1) confirm the independent paleopolyploidization in *Cleome*; 2) assess whether *Cleome* shares the *At*-β duplication with *Arabidopsis*, which is lacking in *Carica* within the Brassicales; and 3) analyze the rates of molecular evolution of the aforementioned Rosid taxa to determine if the herbaceous annual taxa *Arabidopsis* and *Cleome* have a faster rate of molecular evolution than that of predominantly woody perennial lineages. Answering these questions is necessary to further understand the dynamics of genome evolution within one of the most comprehensively surveyed plant genomic systems, the Rosids, and to further resolve the complex genomic history of the *At*-α, *At*-β, and *At*-γ polyploid events of *Arabidopsis*.

Materials and Methods

Plant Material

Cleome spinosa (ES1046; Spinnenpflanze) seeds were obtained from Kiepenkerl. Seeds were sown on a Mini-Tray:vermiculite (3:1) soil mix (Einheitserdenwerk) and cold stratified for 5 days at 4 °C. Afterward, plants were moved to ventilated growth rooms with constant airflow and 40% humidity at 24 °C. Plants were grown at a distance of 30 cm from fluorescent light banks with four bulbs of cool white and four bulbs of wide-spectrum lights at a 14-h light/10-h dark photoperiod. Grow domes were removed after 5 days under lights and plants were fertilized twice with 1 ml of Scotts Peters Professional Peat Lite Special 20N:10P:20K with trace elements and 2 l water per flat, added to the bottom of the tray. Approximately 20 days after germination when plants had developed four true leaves they were transferred to individual pots (15 cm^2) and were grown for up to 3 months under strict light, temperature, and humidity control. Fully emerged vegetative leaf tissue was harvested from plants of different ages (both flowering and nonflowering), with several leaves for each plant and pooled for RNA isolation.

RNA Isolation and cDNA Library Preparation

Isolated plant tissues were immediately submersed in liquid nitrogen and stored at −80 °C. TRIzol Reagent (Invitrogen) was used to isolate the RNA according to the manufacturer's protocol. The RNA was precipitated overnight at −20 °C, and the dried pellet was dissolved in 90 μl RNA Storage Solution (Ambion). Any remaining genomic DNA contamination was removed by DNAse treatment (TURBO DNAse, Ambion). The DNAse enzyme

was removed, and the RNA was further purified by using the RNeasy MinElute Clean up Kit (Qiagen) following the manufacturer's protocol and eluted in 20 μl of RNA Storage Solution (Ambion). Poly(A)+ messenger RNA (mRNA) was purified by binding to an oligo d(T) column (RNA Purist, Ambion). RNA integrity and quantity were verified on an Agilent 2100 Bioanalyzer using RNA Nano chips (Agilent Technologies). RNA quantity was determined on a Nanodrop ND-1000 spectrophotometer.

Full-length–enriched, normalized cDNA libraries were generated using a combination of the SMART cDNA library construction kit (Clontech) and the Trimmer-Direct cDNA normalization kit (Evrogen), generally following the manufacturer's protocol but with several important modifications. In brief, 2 μg of poly(A)+ mRNA was used for the cDNA library generated and reverse transcription was performed with a mixture of several reverse transcription enzymes (ArrayScript, Ambion; BioScript, Bioline; PrimeScript, TaKaRa; SuperScript II, Invitrogen) for 1 h at 42 °C and 90 min at 50 °C. cDNA size fractionation was performed with SizeSep 400 spun columns (GE Healthcare) that resulted in a cutoff at ~200 bp. The full-length–enriched cDNAs were cut with *Sfi*I and ligated to pDNR-Lib plasmid (Clontech). Ligations were transformed into *Escherichia coli* ELECTROMAX DH5α-E electro-competent cells (Invitrogen).

Generation of EST Databases

Plasmid minipreparation from bacterial colonies grown in 96 deep-well plates was performed using the 96-well robot plasmid isolation kit (Eppendorf) on a Tecan Evo Freedom 150 robotic platform (Tecan). Single-pass sequencing of the 5′ termini of cDNA libraries was carried out on an ABI 3730 xl automatic DNA sequencer (PE Applied Biosystems). Vector clipping, quality trimming, and sequence assembly using stringent conditions were conducted with the Lasergene software package (DNAStar Inc.). Blast searches were conducted on a local server using the National Center for Biotechnology Information (NCBI) Blastall program.

Bioinformatic Analyses

Newly generated *C. spinosa* ESTs and genomic data from related Rosid taxa were analyzed using the bioinformatic pipelines described in Barker et al. (2008). Coding sequence (cds) collections for *Arabidopsis thaliana* (TAIR7 CDS, http://www.arabidopsis.org/), *Populus trichocarpa* (v1.1, http://genome.jgi-psf.org/Phypa1_1/Phypa1_1.home.html), and *Vitis vinifera* (v1, http://www.genoscope.cns.fr/spip/Vitis-vinifera-whole-genome.html) were downloaded from their respective databases. EST collections of *Gossypium hirsutum* and *Carica papaya* were downloaded from GenBank. Prior to assembly of EST reads, vector and low-quality sequences were removed using Seqclean (http://compbio.dfci.harvard.edu/tgi/software/) with the UniVec database (http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html). Contigs were assembled for each EST collection using the program TGICL with default settings (http://compbio.dfci.harvard.edu/tgi/software/) (Quackenbush et al. 2000), and a unigene file containing assembled contigs and singletons was created. Assembled unigenes for *Cleome*, *Gossypium*, and *Carica* are available at http://msbarker.com.

For each annotated cds or assembled unigene collection, duplicate gene pairs were identified and their divergence, in terms of substitutions per synonymous site (K_s), was calculated. Duplicate pairs were identified as sequences that demonstrated 40% sequence identity over at least 300 base pairs from a discontinguous MEGABlast (Zhang et al. 2000; Ma et al. 2002). Reading frames for duplicate pairs were identified by comparison to available plant protein sequences. Each duplicated gene was searched against all plant proteins available on GenBank (Wheeler et al. 2008) using BlastX (Altschul et al. 1997). Best-hit proteins were paired with each gene at a minimum cutoff of 30% sequence identity over at least 150 sites. Genes that did not have a best-hit protein at this level were removed before further analyses. To determine reading frame and generate estimated amino acid sequences, each gene was aligned against its best-hit protein by Genewise 2.2.2 (Birney et al. 1996). Using the highest scoring Genewise DNA–protein alignments, custom Perl scripts were used to remove stop and "N"-containing codons and produce estimated amino acid sequences for each gene. Amino acid sequences for each duplicate pair were then aligned using MUSCLE 3.6 (Edgar 2004). The aligned amino acids were subsequently used to align their corresponding DNA sequences using RevTrans 1.4 (Wernersson and Pedersen 2003). K_s values for each duplicate pair were calculated using the maximum likelihood method implemented in codeml of the PAML package (Yang 1997) under the F3x4 model (Goldman and Yang 1994).

Further cleaning of the data set was conducted to remove duplication events that could bias the results. All duplicate pairs containing identifiable transposable elements were removed from the analysis because duplication resulting from transposition may obscure a signal from paleopolyploidy. To reduce the possibility that identical genes are represented in the data set, but missed by the TGICL clustering because of alternative splicing, all K_s values from one member of a duplicate pair with $K_s = 0$ were removed. Further, to reduce the multiplicative effects of multicopy gene families on K_s values, phylogenies for each gene family were constructed by single linkage clustering (Blanc and Wolfe 2004), and node K_s values were calculated. Node K_s values <3 were used in subsequent analyses.

To identify significant features in the age distribution three statistical tests were employed (K-S goodness of fit tests, SiZer, and mixture models). A bootstrapped K-S goodness of fit test was used (Cui et al. 2006) to assess if the overall age distributions deviated from a simulated null of no paleopolyploidizations. Taxa that significantly deviated from the null were then analyzed with SiZer (Chaudhuri and Marron 1999) and EMMIX to identify significant features in our age distributions. SiZer uses changes in the first derivative of a range of kernel density estimates to find significant slope increases or decreases ($\alpha = 0.05$), and the combination may be used to identify peaks and their ranges (Chaudhuri and Marron 1999). A mixture model of normal distributions was fit to the age distribution data by

maximum likelihood using the EMMIX package (Mclachlan et al. 1999). Peaks produced by paleopolyploidy are expected to be approximately Gaussian (Blanc and Wolfe 2004; Schlueter et al. 2004), and this mixture model test identifies the number of normal distributions and their positions that could produce the observed age distributions. For the mixture model analyses, 1–10 normal distributions were fitted to the data with 1,000 random starts and 100 k-mean starts. The Bayesian information criterion was used to select the best model fit to the data because it has a stronger penalty for additional parameters than the Akaike information criterion.

Age distributions from lineages as phylogenetically diverse as lineages of the Brassicales are not directly comparable because of molecular evolutionary rate variation among nuclear genomes. To account for this rate heterogeneity, K_s values for each lineage were corrected using relative rate corrections based on K_s branch length ratios. A representative of each Brassicales lineage with genomic data was included along with three outgroups (*Gossypium*, *Populus*, and *Vitis*) to calculate K_s branch lengths of orthologs across a constrained topology in PAML. Two hundred and seventy putative orthologs with at least 300-bp alignment overlap among these taxa by reciprocal best blast hits (supplementary table S1, Supplementary Material online). Using these orthologs, K_s branch lengths were calculated for each gene in the Brassicales in-group across a constrained topology based on Hall et al. (2004) and Stevens (2008). Using this topology, the ratios of branch lengths for *Cleome*, *Arabidopsis*, and *Carica* versus *Gossypium* were calculated for each gene. The mean ratio over all 270 orthologs for each lineage was applied as a relative rate correction to the K_s values for their respective taxa, and one-sided Mann–Whitney *U* tests of the rate-corrected duplication distributions and the distribution of ortholog divergences were used to assess if duplications occurred after lineage divergence. The rate-corrected values of putatively shared duplications were further compared with analyses of variance (ANOVAs) to determine if the rate-corrected means are significantly different from each other.

Phylogenetic Analysis of Potential *At*-β Duplicated Loci

Based on the analysis of K_s distributions (above), potential *Cleome* EST paleolog sets with a K_s divergence of between 1.2 and 1.9 were further analyzed to address if the *At*-β duplication is shared between *Cleome* and *Arabidopsis*. First, homologous sequences from *Arabidopsis* were checked for maintained *At*-β (and *At*-α) paleologs according to previous studies by Blanc et al. (2003) and Bowers et al. (2003). Only those that were known to have *At*-β paleologs in *Arabidopsis* (based on stretches of at least seven collinear *At*-β paleologs) were further analyzed. Analyses were further focused on those sequences that were not parts of large multi-gene families (e.g., with more than 10 gene members in *Arabidopsis*) and preferentially those that also had at least one set of *At*-α paleologs in *Arabidopsis*. Secondly, homologous loci from other taxa were identified using both Blast analysis and available syntenic information provided by the Plant Genome Duplication Database (PGML, http://chibba.agtec.uga.edu/duplication/) and Phytozome (http://www.phytozome.net/). Note, that *At*-γ paleologs were also included when information was available. Third, the deduced proteins for each of these paleolog sets were aligned using either Clustal and/or using the sequence alignment server at MAFFT (v. 6, http://align.bmr.kyushu-u.ac.jp/mafft/online/server/). The corresponding DNA alignment was manually inspected and adjusted. Well-aligned regions were then used for phylogenetic analysis. Sequences were first analyzed using Modeltest (Posada and Crandall 1998) and then a phylogeny using the appropriate model was generated using PhymL 3.0 (Guindon et al. 2005).

Data Deposition

All sequences are deposited under GenBank accessions GR930901–GR935577.

Results

Our analysis of gene duplications in the Brassicales reveals a history of multiple large-scale genome duplications. We obtained 4,764 high-quality 5′-end sequenced clones from our *Cleome*-normalized and size-selected cDNA library. After cleaning and assembly, these sequences yielded 4,677 unigenes. Of the unigenes, 542 showed evidence of being duplicated (11.7% of the unigenes). Because we sequenced a normalized cDNA library, the number of sequenced clones is very similar to the number of assembled unigenes and for most of the more recent duplications (e.g., with $K_s < 0.3$) we likely sequenced a single member. Thus, the overall percentage of unigenes detected in duplicate is low for *Cleome*. Importantly, the *Cleome* data set, in terms of overall unigene number and the number of duplicated unigenes, is comparable to or larger than previous EST data sets used to successfully infer paleopolyploidy (Cui et al. 2006; Barker et al. 2008). For *A. thaliana*, 63.3% of genes were duplicated as well as 39.1% of genes in *Carica*. Histograms of the age of gene duplication events, as inferred by K_s from our gene family phylogenies, demonstrated peaks consistent with paleopolyploidy in the ancestry of *Cleome*, as well as confirming those already reported for *Arabidopsis* and *Carica* (fig. 1). Consistent with significant peaks in the histograms, a K-S goodness of fit test indicates that the K_s distributions also deviated significantly ($P < 0.0001$) from a null model of constant duplicate gene birth and death. SiZer analyses identified significant peaks ($P < 0.05$) in the age distributions of *Cleome* and *Arabidopsis* that corresponded to peaks in the histograms (Supplementary Results, Supplementary Material online).

Mixture model analyses recover multiple peaks in the three Brassicales taxa, some of which correspond to previously recognized paleopolyploidizations (table 1). For *Arabidopsis*, mixture models identified three distributions that overlapped with histogram and SiZer peaks that are consistent with the previously described *At*-α (median $K_s = 0.70$), *At*-β (median $K_s = 1.73$), and *At*-γ (median $K_s = 2.67$) paleopolyploidizations (fig. 1*A*). In contrast, we only found a single peak in *Carica* (median $K_s = 1.22$), consistent with

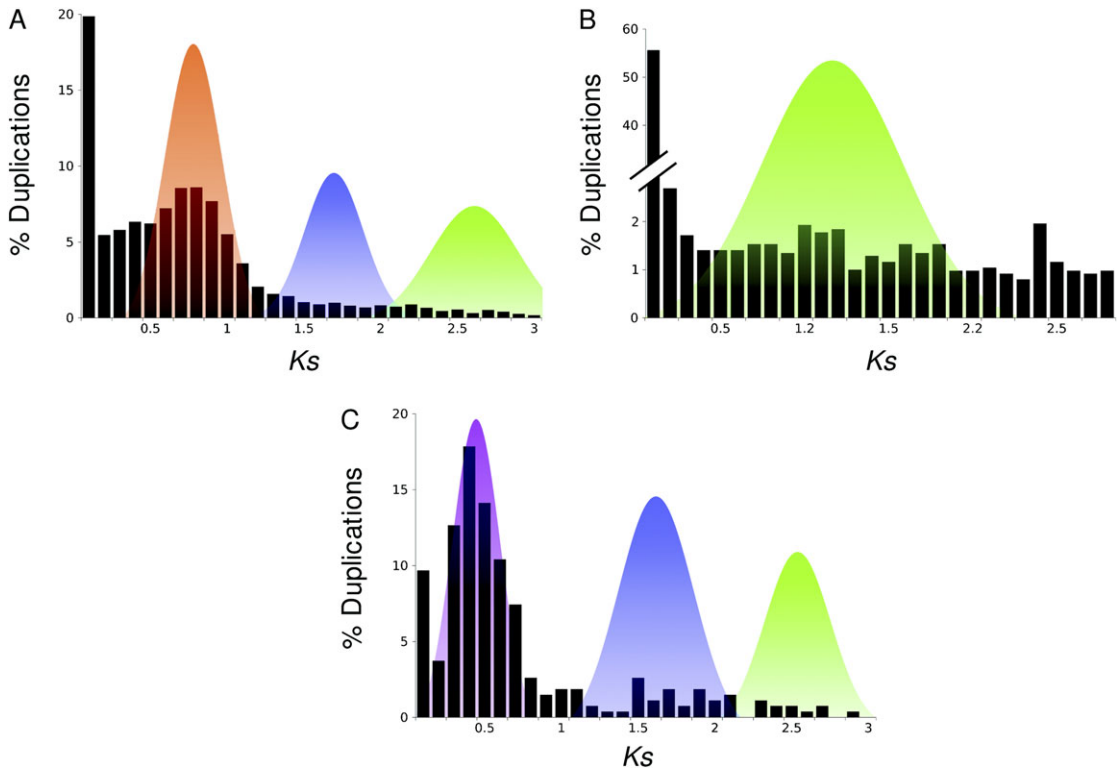

FIG. 1.—Histograms of Brassicales gene duplication ages with mixture model fits. (*A*) *A. thaliana*, Brassicaceae. (*B*) *C. papaya*, Caricaceae. (*C*) *C. spinosa*, Cleomaceae. Plots of normal distributions were fitted from mixture model analyses; the orange plot represents *At*-α, the blue plot represents *At*-β, green plot represents *At*-γ, and the purple plot represents *Cs*-α.

recent reports (Ming et al. 2008; Tang, Wang et al. 2008) that it experienced only the most ancient *At*-γ event (fig. 1*B*). Similar to *Arabidopsis*, we found evidence of three peaks in *Cleome*, with medians at K_s = 0.41, 1.68, and 2.53 (fig. 1*C*). For each species, the deltaBIC scores for the selected models were greater than 2 and suggest that models including paleopolyploidization explain the data significantly better than other models without such large genome duplications (table 1).

To resolve the number and phylogenetic placement of duplications, we used our mean phylogeny of 270 Brassicales nuclear ortholog sets (fig. 2; tables 1 and 2; and supplementary tables S1 and S2, Supplementary Material online). Surprisingly, we observe that the mean ratios of K_s branch lengths for *Cleome* and *Arabidopsis* versus the *Gossypium* reference are nearly identical at 2.05 and 2.06, respectively. *Carica* is evolving at less than half the rate of these two derived Brassicales lineages, with a mean branch length ratio of 0.98 versus *Gossypium*. Taking into account this rate heterogeneity, we calculate a mean rate-corrected divergence between *Arabidopsis* and *Cleome* of K_s = 0.41, whereas *Carica* and these two lineages diverged at K_s = 0.89. After correcting the duplication peak medians with the appropriate ratio, we find that the youngest peak in *Arabidopsis*, *At*-α, is centered at K_s = 0.34, after the divergence of the Brassicaceae and Cleomaceae. Similarly, the youngest peak in *Cleome* is centered at K_s = 0.20 after rate correction. Mann–Whitney *U* tests indicate that the medians of each polyploidization are located after the median ortholog divergence for *Arabidopsis* and *Cleome* (supplementary table S2, Supplementary Material online). Thus, *At*-α is restricted to the Brassicaceae, whereas the Cleomaceae has experienced an independent genome duplication, named here as *Cs*-α, consistent with the analyses of Schranz and Mitchell-Olds (2006).

Analyses of rate heterogeneity and divergence also indicate that *At*-β is shared by the Brassicaceae and Cleomaceae, but not *Carica* (fig. 2, supplementary tables S1 and S2, Supplementary Material online). The rate-corrected *At*-β peak median in *Arabidopsis* is located at K_s = 0.84. In *Cleome*, a similar peak is observed at a rate-corrected K_s = 0.82 and is not significantly different from the *Arabidopsis* peak (ANOVA *P* value = 0.91). Importantly, these peaks occur after the divergence of *Carica* from these other two Brassicales at K_s = 0.89

Table 1
Summary of Mixture Model Distributions Inferred to Represent Paleopolyploidizations across Three Brassicales Taxa

Family	Genus	Mixture Median (K_s)[a]	% of Data	△BIC (w/o *At*-γ)	△BIC (w/o *At*-β and *At*-γ)
Brassicaceae	Arabidopsis	0.69795	49	133.33	1667.13
		1.7324	14.6		
		2.6762	6.1		
Cleomaceae	*Cleome*	0.41307	72	24.81	295.29
		1.6791	15.3		
		2.5267	2.3		
Caricaceae	*Carica*	1.2209	24.4	6106.93	

BIC, Bayesian information criterion.

[a] Complete mixture model distributions available in supplementary table S1 (Supplementary Material online).

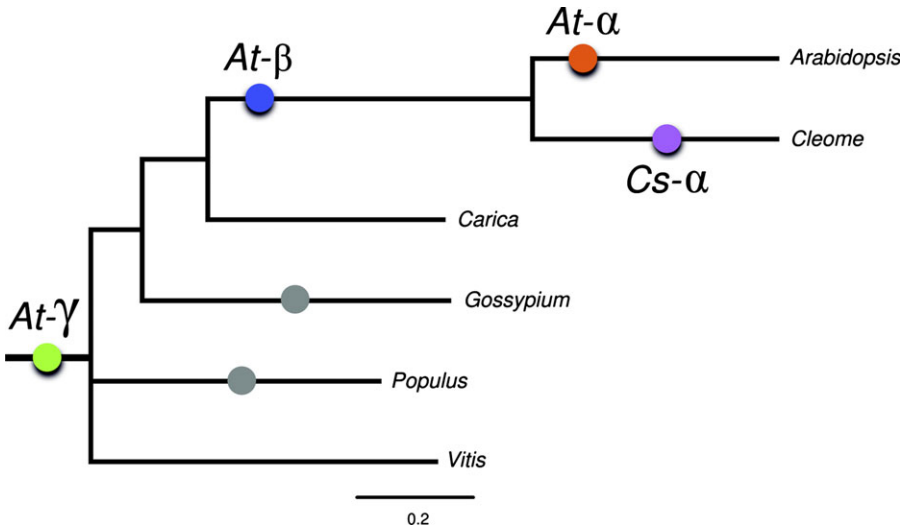

FIG. 2.—Phylogeny of Brassicales taxa and related Rosid outgroups displaying inferred paleopolyploidizations. Branch lengths are mean K_s values from 270 nuclear ortholog sets (supplementary table S1, Supplementary Material online). Colored dots indicate inferred paleopolyploidizations placed in relation to lineage divergence based on our rate corrections. Gray dots represent paleopolyploidizations inferred in previous analyses (Blanc and Wolfe 2004; Sterk et al. 2005).

(supplementary table S2, Supplementary Material online), consistent with previous synteny analyses indicating that *Carica* lacks *At*-β (Tang, Bowers et al. 2008). Similarly, the oldest peak that corresponds to *At*-γ is located at a rate-corrected median of $K_s = 1.29$ in *Arabidopsis*, 1.23 in *Cleome*, and 1.24 in *Carica* (ANOVA P value = 0.54). These peaks are all located before the divergence of the Brassicales from other Rosids, in this case *Gossypium*, at $K_s = 1.10$.

To further confirm the phylogenetic circumscription of *At*-β, we scrutinized *Cleome* gene families for duplicate gene sets that may be derived from the *At*-β duplication (K_s divergence at 1.2–1.9). We identified 32 *Cleome* paleolog sets in this K_s divergence range. From these 32 sets, we focused on 11 sets of *Cleome* duplicates for which there was strong evidence of maintained *At*-β duplicate pairs in *Arabidopsis* based on synteny and that were not a member of large multi-gene families. Of these 11 potential *At*-β–derived gene families, seven families also had a set of maintained *At*-α–derived duplicates. Such nested retained duplicates are usually dosage-sensitive genes from small gene families (and therefore have more protein–protein interactions) (Aury et al. 2006; Freeling 2009). We queried various databases (including Phytozome, PGML, and NCBI) to identify homologous loci from other species to include in our seven "*At*-β sets" for further phylogenetic analysis. Consistent with the duplicate gene age distributions, phylogenetic analyses of these gene families demonstrate that *At*-β is shared by *Arabidopsis* and *Cleome* and that *At*-α is restricted to the Brassicaceae (fig. 3, supplementary table S3, Supplementary Material online).

Table 2
Rate-Corrected Mixture Model Medians of Brassicales Paleopolyploidizations

		Rate-Corrected Paleopolyploidizations (K_s)			
Genus	Relative Rate (% K_s)[a]	*At*-α	*Cs*-α	*At*-β	*At*-γ
Arabidopsis	+206	0.34		0.84	1.29
Cleome	+205		0.20	0.82	1.23
Carica	−2.4				1.24

[a] Percent difference relative to *Gossypium*, based on 270 nuclear gene ortholog phylogenies (supplementary table S1, Supplementary Material online).

Discussion

Our results confirm that the Cleomaceae has a paleopolyploidization independent of the *At*-α duplication in the Brassicaceae. Schranz and Mitchell-Olds (2006) had previously shown that *Cleome* had an ancient genome polyploidization based on analyses of BAC sequences homologous to duplicated blocks in *Arabidopsis*. Phylogenetic reconstructions of selected gene families in these BACs suggested that the *Cleome* polyploidization was likely independent of the well-studied *At*-α duplication of the Brassicaceae. Our analyses lend further support to this interpretation and find that the *Cleome* polyploidization, *Cs*-α, occurred well after the divergence of the Brassicaceae and Cleomaceae. The phylogenetic reconstructions of Schranz and Mitchell-Olds (2006) also found evidence that the *Cleome* polyploidization was actually a triplication (ancient hexaploidy). The analyses presented here do not provide further resolution to this question, and more complete gene family sequencing and BAC analyses are needed to confirm this hypothesis.

Our results also confirm that *At*-α, perhaps the most well studied plant genome duplication, is indeed restricted to the Brassicaceae. Previously, the phylogenetic circumscription of *At*-α was largely unresolved, with some studies placing it as old as the origin of the Brassicales (Ku et al. 2000; Lynch and Connery 2000; Bowers et al. 2003), near the origin of the Brassicaceae (Blanc et al. 2003; Ermolaeva et al. 2003), and following the divergence of most Brassicaceae from the genus *Aethionema* (Galloway et al. 1998; Schranz and Mitchell-Olds 2006). Most analyses placed *At*-α using absolute dating approaches, and the placement has

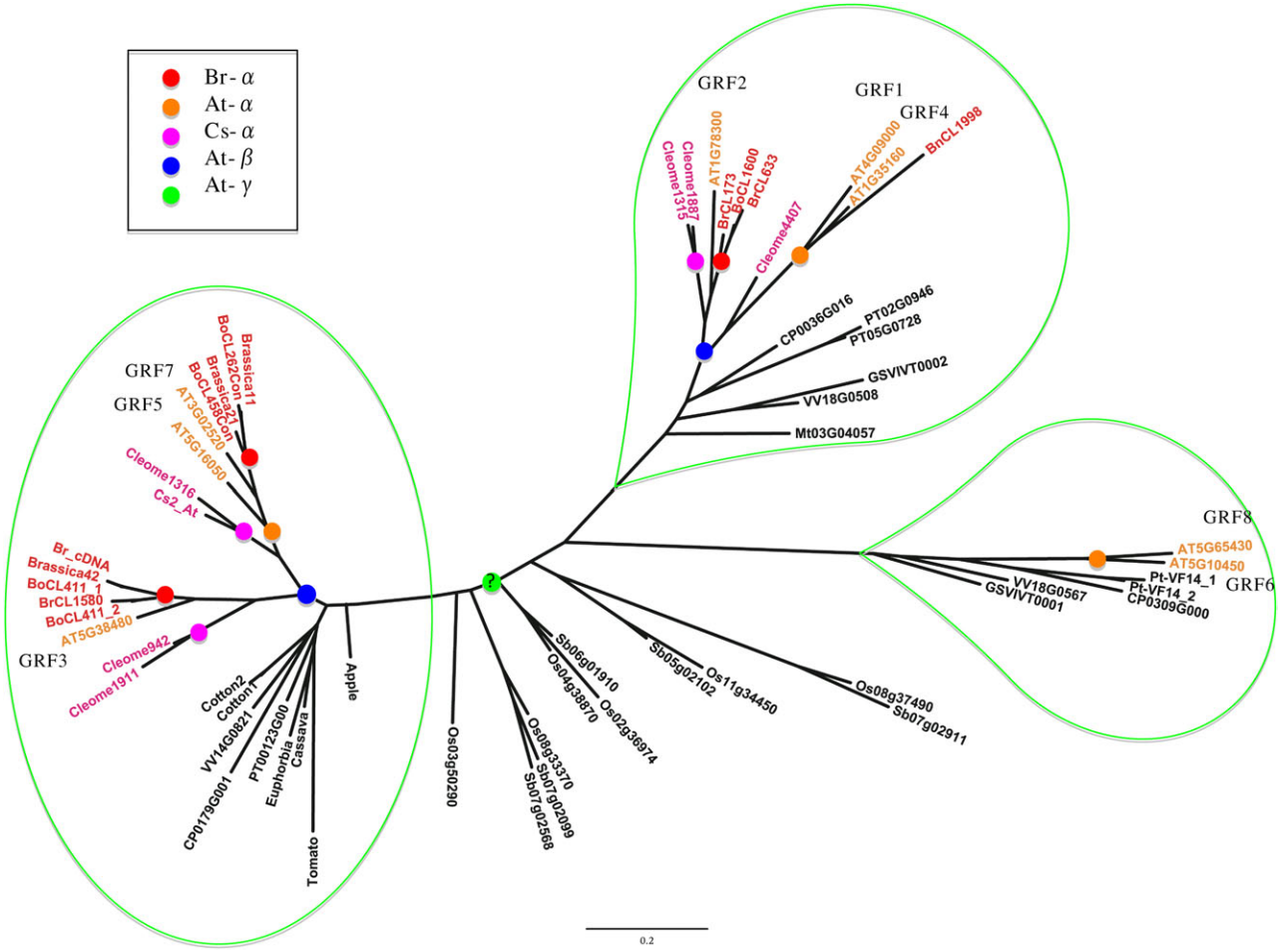

FIG. 3.—Phylogeny of the General Regulatory Factor (GRF or 14-3-3) gene family with nodes highlighted that support inferred paleopolyploidy events. Sequences from key taxa used to infer polypoloidizations are labeled: *Brassica* (red), *Arabidopsis* (orange), and *Cleome* (pink). Nodes supporting particular paleopolyploidy events (i.e., clades containing paralogous loci) are labeled with colored circles: *Br*-α in red (*Brassica* genome triplication), *At*-α in orange (Brassicaceae genome duplication), *Cs*-α in pink (*Cleome* genome triplication), *At*-β in blue (core-Brassicales genome duplication), and the potential placement of the *At*-γ is shown in green (Rosid or Angiosperm genome triplication). For the analysis, EST sequences from several species (sequences labeled by genus or common names) as well as from the following genomic sequences: Os (*Oryza sativa*), Sb (*Sorghum bicolor*), Mt (*Medicago truncatula*), Vv (*V. vinifera*), Pt (*P. trichocarpa*), Cp (*C. papaya*), and At (*A. thaliana*). The eight members of the *Arabidopsis* gene family used in the analysis are additionally labeled by their gene family names: GRF1–GRF8.

largely varied with the dating method or calibration. Our approach avoids the entire dating problem by placing the duplication in relation to the divergence between the Brassicaceae and its sister family and conclusively shows that *At*-α is restricted to at least the Brassicaceae. However, we can estimate the age of *At*-α using synonymous substitutions rates calibrated for the Brassicaceae because our results demonstrate its restricted to the family. Based on the calibration of Koch et al. (2000) and the median K_S of the *At*-α peak, we estimate that the *At*-α duplication occurred approximately 23.3 Ma, placing it near the diversification of the derived clades of the Brassicaceae. Because our analyses show that the substitution rate of the Cleomaceae is nearly identical to that of the Brassicaceae, we can similarly use the Koch et al. (2000) calibration to date *Cs*-α. Using this approach, we estimate that the putative triplication *Cs*-α occurred roughly 13.7 Ma.

We also resolve some of the mystery surrounding the existence and placement of the *At*-β duplication. Although Bowers et al. (2003) and Tang, Bowers et al. (2008) resolved three genome duplications in *Arabidopsis*, other analyses have been less clear on the number. The youngest duplication, *At*-α, has consistently been resolved, but *At*-β has not always been clearly distinguished from the older *At*-γ duplication (Vision et al. 2000; Blanc et al. 2003; Blanc and Wolfe 2004). Contributing to questions surrounding the existence of *At*-β was its absence in the *Carica* genome (Ming et al. 2008), a member of the Brassicales. This was a surprising result because *At*-β was thought to have occurred well before the origin of the Brassicales (Bowers et al. 2003). Our observation of *At*-β in only *Cleome* and *Arabidopsis* is consistent with these results and confirms the existence of *At*-β by providing evidence of the duplication in a second lineage beyond *Arabidopsis*. Further, our relative phylogenetic placement indicates that *At*-β occurred after the split of *Carica* from the remaining Brassicales, explaining its absence from the *Carica* genome. Phylogenetic reconstructions of the order Brassicales show that shortly after the divergence of the family Caricaceae there was a radiation of several families (including

Brassicaceae and Cleomaceae) making up the "core-Brassicales" (Hall et al. 2004). We hypothesize that the *At*-β event may correlate with the diversification of the core-Brassicales clade (including the families Brassicaceae, Cleomaceae, Capparaceae, Resedaceae, Gyrostemonaceae, Pentadiplandraceae, Tovariaceae, and Emblingiaceae). Previous attempts to place *At*-β were clearly confounded by nuclear genome rate heterogeneity (Tang, Wang et al. 2008), and our rate-standardized analyses reconcile the previously conflicting data on *At*-β.

Given the substantial nuclear genome substitution rate heterogeneity observed in previous analyses (Barker et al. 2008; Tang, Wang et al. 2008), we were surprised to find no rate heterogeneity between the Cleomaceae and Brassicaceae. We did, however, find that these two families are evolving more than twice as fast as other Rosid lineages, including *Carica* and *Gossypium*. One potentially significant difference between these two groups is that the Cleomaceae and Brassicaceae are predominantly herbaceous annuals, whereas the remaining lineages are composed mostly of perennials (Stevens 2008). Increased substitution rates in the nuclear genomes of annuals relative to perennials is a consistent observation across a wide variety of plant lineages (Gaut and Clegg 1991; Gaut et al. 1996; Koch et al. 2000), and Smith and Donoghue (2008) recently demonstrated that life histories are highly correlated with rates of molecular evolution in plants. Despite these observations, Tang, Wang et al. (2008) suggested that life history is not sufficient to explain the rate heterogeneity observed among the Brassicales and other Rosid taxa. We believe that life history may explain much of the observed rate heterogeneity. Across our 270 nuclear ortholog phylogenies, we find that the *Arabidopsis* nuclear genome is evolving at a mean of 2.11 times the rate of the *Carica* genome, similar to the synonymous substitution rate differences observed by Smith and Donoghue (2008) among related woody and herbaceous plants. Further, assuming that the synonymous substitution rate of *Arabidopsis* is 1.5×10^{-8} (Koch et al. 2000), the estimated synonymous substitution rate of *Carica* is approximately 7.5×10^{-9}. Considering that the average synonymous substitution rate for plants is 6.1×10^{-9} (Lynch and Connery 2000), with rates of $0.70–1.31 \times 10^{-9}$ in long-lived perennials such as pines (Ann et al. 2007), the rate heterogeneity observed in the Brassicales and related Rosids is well within the range of reported synonymous substitution rates in plants. Because life history varies considerably across the plant phylogeny, future work on paleopolyploidy will need to account for rate heterogeneity to place duplications more accurately in phylogenetic context as presented here and in Barker et al. (2008) or with reconciliation approaches as in Pfeil et al. (2005).

Although we have further elucidated the position and number of duplications in the ancestry of *Arabidopsis* and related Brassicales, substantial work is still needed for complete resolution. Genomic data from additional lineages of the Brassicales (e.g., from the Resedaceae in the core-Brassicales and Limnanthaceae not in the core-Brassicales) are needed to more precisely place *At*-β and to address if it correlates with the diversification of the core-Brassicales clade (Hall et al. 2004). Similarly, additional genomic sampling within the Brassicaceae and Cleomaceae is required to place the *At*-α and *Cs*-α paleopolyploidizations more precisely. Such sampling will likely prove fruitful for a wide range of genomic studies, particularly to further understanding of the relationship between paleopolyploidy and diversification in these large angiosperm families.

Supplementary Material

Supplementary results (tables S1–S3) are available at *Genome Biology and Evolution* online (http://www.oxfordjournals.org/our_journals/gbe/).

Unigene assemblies for all species are available at http://msbarker.com.

Funding

National Institutes of Health Training (T32 GM007757 to M.S.B.).

Acknowledgments

Comments from P. Edger, G. J. Gastony, M. W. Hahn, J. C. Pires, and L. H. Rieseberg improved the manuscript.

Literature Cited

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Ann W, Syring J, Gernandt DS, Liston A, Cronn R. 2007. Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations for *Pinus*. Mol Biol Evol. 24:90–101.

Aury JM, et al. 2006. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature. 444:171–178.

Barker MS, 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.

Birney E, Thompson J, Gibson T. 1996. PairWise and SearchWise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. Nucleic Acids Res. 24:2730–2739.

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

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

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

Chapman BA, Bowers JE, Feltus FA, Paterson AH. 2006. Buffering of crucial functions by paleologous duplicated genes may contribute cyclicality to angiosperm genome duplication. Proc Natl Acad Sci USA. 103:2730–2735.

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

Cui L, et al. 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.
Doyle JJ, et al. 2008. Evolutionary genetics of genome merger and doubling in plants. Annu Rev Genet. 42:443–461.
Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.
Ermolaeva MD, Wu M, Eisen JA, Salzberg SL. 2003. The age of the *Arabidopsis thaliana* genome duplication. Plant Mol Biol. 51:859–866.
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.
Gaeta RT, Pires JC, Iniguez-Luy F, Leon E, Osborn TC. 2007. Genomic changes in resynthesized *Brassica napus* and their effect on gene expression and phenotype. Plant Cell. 19:3403–3417.
Galloway GL, Malmberg RL, Price RA. 1998. Phylogenetic utility of the nuclear gene arginine decarboxylase: an example from the Brassicaceae. Mol Biol Evol. 15:1312–1320.
Gaut BS, Clegg MT. 1991. Molecular evolution of alcohol dehydrogenase 1 in members of the grass family. Proc Natl Acad Sci USA. 88:2060–2064.
Gaut BS, Morton BR, McCaig BC, Clegg MT. 1996. Substitution rate comparisons between grasses and palms: synonymous rate differences in the nuclear gene Adh parallel rate differences at the plastid gene rbcL. Proc Natl Acad Sci USA. 93:10274–10279.
Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 11:725–736.
Guindon S, Lethiec F, Duroux P, Gascuel O. 2005. PHYML Online—a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res. 33:W557–W559.
Hall JC, Iltis HH, Sytsma KJ. 2004. Molecular phylogenetics of core brassicales, placement of orphan genera *Emblingia*, *Forchhammeria*, *Tirania*, and character evolution. Syst Bot. 29:654–669.
Jaillon O, et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature. 449:463–467.
Koch MA, Haubold B, Mitchell-Olds T. 2000. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). Mol Biol Evol. 17:1483–1498.
Ku HM, Vision T, Liu J, Tanksley SD. 2000. Comparing sequenced segments of the tomato and *Arabidopsis* genomes: large-scale duplication followed by selective gene loss creates a network of synteny. Proc Natl Acad Sci USA. 97: 9121–9126.
Lyons E, et al. 2008. Finding and comparing syntenic regions among Arabidopsis and the outgroups Papaya, Poplar, and Grape: CoGe with Rosids. Plant Physiol. 148:1772–1781.
Lynch M, Connery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science. 290:1151–1155.
Ma B, Tromp J, Li M. 2002. PatternHunter: faster and more sensitive homology search. Bioinformatics. 18:440–445.
Maere S, et al. 2005. Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci USA. 102:5454–5459.
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.
Ming R, et al. 2008. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature. 452:991–997.
Pfeil BE, Schlueter JA, Shoemaker RC, Doyle JJ. 2005. Placing paleopolyploidy in relation to taxon divergence: a phylogenetic analysis in legumes using 39 gene families. Syst Biol. 54:441–454.
Posada D, Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics. 14:817–818.
Quackenbush J, Liang F, Holt I, Pertea G, Upton J. 2000. The TIGR gene indices: reconstruction and representation of expressed gene sequences. Nucleic Acids Res. 28:141–145.
Schlueter JA, et al. 2004. Mining EST databases to resolve evolutionary events in major crop species. Genome. 47:868–876.
Schranz ME, Mitchell-Olds T. 2006. Independent ancient polyploidy events in the sister families Brassicaceae and Cleomaceae. Plant Cell. 18:1152–1165.
Smith SA, Donoghue MJ. 2008. Rates of molecular evolution are linked to life history in flowering plants. Science. 322: 86–89.
Soltis DE, et al. 2009. Polyploidy and angiosperm diversification. Am J Bot. 96:336–348.
Sterck L, et al. 2005. EST data suggest that poplar is an ancient polyploid. New Phytol. 167:165–170.
Stevens PF. 2008. Angiosperm Phylogeny Website [Internet]. St. Louis (MO): University of Missouri and Missouri Botanical Garden; Version 9 released 28 May 2008 [cited 1 Jun 2008]. Available from: http://www.mobot.org/MOBOT/research/APweb/.
Tang H, Bowers JE, et al. 2008. Synteny and collinearity in plant genomes. Science. 320:486–488.
Tang HB, Wang XY, et al. 2008. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. Genome Res. 18:1944–1954.
Town CD, et al. 2006. Comparative genomics of *Brassica oleracea* and *Arabidopsis thaliana* reveal gene loss, fragmentation, and dispersal after polyploidy. Plant Cell. 18:1348–1359.
Vision TJ, Brown DG, Tanksley SD. 2000. The origins of genomic duplications in *Arabidopsis*. Science. 290: 2114–2117.
Wernersson R, Pedersen AG. 2003. RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res. 31:3537–3539.
Wheeler DL, et al. 2008. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 36:D13–D21.
Wolfe K. 2000. Robustness–it's not where you think it is. Nat Genet. 25:3–4.
Wood TE, et al. 2009. The frequency of polyploid speciation in vascular plants. Proc Natl Acad Sci USA. 106:13875–13879.
Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13:555–556.
Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences. J Comput Biol. 7:203–214.

Michael Purugganan, Associate Editor

Accepted October 1, 2009