Annals of Botany **106**: 497–504, 2010
doi:10.1093/aob/mcq129, available online at www.aob.oxfordjournals.org

ANNALS OF BOTANY
Founded 1887

PART OF A HIGHLIGHT ON GENES IN EVOLUTION

Ancient genome duplications during the evolution of kiwifruit (*Actinidia*) and related Ericales

Tao Shi[1], Hongwen Huang[1,2] and Michael S. Barker[2,3,4,*]

[1]*Key Laboratory of Plant Germplasm Enhancement and Speciality Agriculture, Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan 430074, Hubei, China,* [2]*South China Botanical Garden/South China Institute of Botany, Chinese Academy of Sciences, Guangzhou, Guangdong, China,* [3]*The Biodiversity Research Centre, University of British Columbia, Vancouver, BC, Canada V6T 1Z4 and* [4]*Department of Biology, Indiana University, Bloomington, IN 47405, USA*
** For correspondence. E-mail msbarker@biodiversity.ubc.ca*

Received: 24 March 2010 Returned for revision: 20 April 2010 Accepted: 20 May 2010 Published electronically: 24 June 2010

• *Background and Aims* To assess the number and phylogenetic distribution of large-scale genome duplications in the ancestry of *Actinidia*, publicly available expressed sequenced tags (ESTs) for members of the Actinidiaceae and related Ericales, including tea (*Camellia sinensis*), were analysed.
• *Methods* Synonymous divergences (K_s) were calculated for all duplications within gene families and examined for evidence of large-scale duplication events. Phylogenetic comparisons for a selection of orthologues among several related species in Ericales and two outgroups permitted placement of duplication events in relation to lineage divergences. Gene ontology (GO) categories were analysed for each whole-genome duplication (WGD) and the whole transcriptome.
• *Key Results* Evidence for three ancient WGDs in *Actinidia* was found. Analyses of paleologue GO categories indicated a different pattern of retained genes for each genome duplication, but a pattern consistent with the dosage-balance hypothesis among all retained paleologues.
• *Conclusions* This study provides evidence for one independent WGD in the ancestry of *Actinidia* (*Ad*-α), a WGD shared by *Actinidia* and *Camellia* (*Ad*-β), and the well-established *At*-γ WGD that occurred prior to the divergence of all taxa examined. More ESTs in other taxa are needed to elucidate which groups in Ericales share the *Ad*-β or *Ad*-α duplications and their impact on diversification.

Key words: Paleopolyploidy, Actinidiaceae, Ericales, *Actinidia*, *Camellia*, kiwi, genome duplication, dosage balance.

INTRODUCTION

The importance of polyploidy, or whole genome duplication, in plant evolution has been long recognized by researchers. Nearly 35 % of flowering plants are of recent polyploid provenance and at least 15 % of angiosperm speciation events are caused by whole genome duplication (Wood *et al.*, 2009). Using recently developed genomic approaches, several ancient genome duplication events have also been inferred during the evolution of flowering plants (Blanc and Wolfe, 2004; Cui *et al.*, 2006; Barker *et al.*, 2008; Soltis *et al.*, 2009). Recent polyploids are easy to detect by changes in chromosome numbers, genome size and gene copy number compared with progenitors, but ancient polyploids, or paleopolyploids, are much harder to identify because diploidization, gene loss and chromosomal rearrangements erode the signal. Despite the obfuscating action of these forces, paleopolyploidy may still be inferred by recognition of homoeologous chromosomes or large bursts of gene duplication (Barker and Wolf, 2010).

Although completely sequenced and assembled nuclear genomes provide the ultimate resource for inferring paleopolyploidy, the identification of peaks of gene duplications in expressed sequenced tags (ESTs) provides an economical method to survey ancient polyploidy. The thousands of ESTs available for many plants provide a useful 'snapshot' of each genome. Large-scale duplication events lead to a punctuated, dramatic increase in the number of duplicated genes (Lynch and Connery, 2000; Blanc and Wolfe, 2004). The resulting excess of paralogues of a particular age produces a peak in the age distribution of duplications across gene families within a genome. This approach has been successfully employed in a variety of plants including wheat (Blanc and Wolfe, 2004), maize (Blanc and Wolfe, 2004; Schlueter *et al.*, 2004), *Solanum* (Schlueter *et al.*, 2004; Blanc and Wolfe, 2004; Cui *et al.*, 2006), *Populus* (Sterck *et al.*, 2005), *Arabidopsis* (Blanc and Wolfe, 2004; Maere *et al.*, 2005; Barker *et al.*, 2009), lettuce (Barker *et al.*, 2008) and sunflower (Barker *et al.*, 2008). Researchers have begun to elucidate the phylogenetic position of paleopolyploidizations in relation to lineage divergence by combining genomic and phylogenetic methods in the Fabaceae (Pfeil *et al.*, 2005), Compositae (Barker *et al.*, 2008) and Brassicales (Bowers *et al.*, 2003; Schranz and Mitchell-Olds, 2006; Tang *et al.*, 2008; Barker *et al.*, 2009).

The genus *Actinidia*, well known as kiwifruit, contains 76 species of climbing plants originating mainly in China (Huang and Ferguson, 2007). Over the past three decades, kiwifruit has developed into an important horticultural cash crop and a fruit industry worldwide (Huang and Ferguson,

© The Author 2010. Published by Oxford University Press on behalf of the Annals of Botany Company. All rights reserved.
For Permissions, please email: journals.permissions@oxfordjournals.org

2007). Recently, a collection of 132 577 ESTs in *Actinidia* were sequenced and publicly released (Crowhurst *et al.*, 2008). Comparative cytological studies among the three extant genera of the Actinidiaceae – *Saurauia* ($x = 13$), *Clematoclethra* ($x = 12$) and *Actinidia* ($x = 29$) – suggest that *Actinidia* is a paleotetraploid derived from an ancestor with $x = 14$ (He *et al.*, 2005). However, no other study has confirmed this hypothesis. Here, publicly available sources of ESTs are used for *Actinidia*, several related genera, such as *Camellia* and *Diospyros* in the Ericales and *Populus* and *Vitis* as outgroups to (*a*) identify ancient genome duplications in *Actinidia* and other Ericales, (*b*) place genome duplications onto the current phylogeny, (*c*) analyse the gene ontology (GO) patterns of retained duplicates from putative whole-genome duplications (WGDs).

MATERIALS AND METHODS

Unigene assembly

EST collections of three *Actinidia* species (*A. chinensis*, 47 380 ESTs; *A. deliciosa*, 57 752 ESTs; *A. eriantha*, 12 648 ESTs), *Camellia sinensis* (10 431 ESTs) and *Diospyros kaki* (9475 ESTs) were downloaded from GenBank. Simulations (Cui *et al.*, 2006) have indicated that species with 10 000 or more ESTs are sufficient for inferring ancient polyploidy, and the analysed *Actinidia* and *Camellia* data are beyond this threshold. Annotated coding sequences (cds) from the whole genome sequences of *Populus trichocarpa* (v1·1, http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html) (Tuskan *et al.*, 2006) and *Vitis vinifera* (v1, http://www.genoscope.cns.fr/spip/Vitis-vinifera-whole-genome.html) (Jaillon *et al.*, 2007) were downloaded from their project sites. For the EST reads, vector and low quality sequences were removed using Seqclean with the UniVec contaminant database (http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html). Contigs were assembled for each EST collection by TGICL with default settings (Quackenbush *et al.*, 2000), and a unigene file containing assembled contigs and singletons was created. Raw reads and assembled unigenes are available at http://msbarker.com and http://biotorrents.net.

Gene family construction and K_s *calculation*

For each annotated cds or assembled unigene collection, gene families were identified and their duplications, in terms of substitutions per synonymous site (K_s), was calculated. Duplicate pairs were identified as sequences that demonstrated 40 % sequence similarity over at least 300 bp from a discontinuous all-against-all MegaBLAST (Zhang *et al.*, 2000; Ma *et al.*, 2002). Pairs of genes 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 calculated. Node K_s values <2 were used in subsequent analyses.

Calculations of the synonymous divergence among gene copies were based upon protein-guided DNA alignments of gene families. Each duplicated gene was searched against all plant proteins available on GenBank (Wheeler *et al.*, 2007) using BLASTX (Altschul *et al.* 1997). Best-hit proteins were paired with each gene at a minimum cutoff of 30 % sequence similarity 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).

Identification of paleopolyploidy

Two statistical tests were employed to identify significant features in the age distribution. 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. Taxa that significantly deviated from the null were then analysed using a mixture model. 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 (Schlueter *et al.*, 2004; Blanc and Wolfe, 2004), and the mixture model identifies the number of normal distributions and their position(s) that best explain the observed age distributions. For the mixture model analyses, one to ten normal distributions were fitted to the data with 1000 random starts and 100 k-mean starts. The Bayesian Information Criterion (BIC) was used to select the best model because it more strongly penalizes increasing the number of model parameters than the Akaike Information Criterion and should be more robust against fitting insignificant distributions.

Phylogenetic placement of ancient duplications and age estimation

To place duplications in relative phylogenetic context and account for substitution rate heterogeneity among members of the Ericales, K_s values for each lineage were corrected using relative rate corrections based on K_s branch length ratios. A representative of each Ericales lineage with genomic data was included along with two outgroups (*Populus* and *Vitis*) to calculate K_s branch lengths of orthologues across a constrained topology in PAML. Putative orthologues were identified as reciprocal best blast hits among these

taxa with at least 300 bp alignment overlap (Table S2 in Supplementary data, available online). Using these orthologues, K_s branch lengths were calculated for each gene in the Ericales ingroup across a constrained topology (Anderberg *et al.*, 2002). For each orthologue phylogeny, the ratios of branch lengths for *Actinidia deliciosa* and *Camellia sinensis* versus *Diospyros kaki* were calculated. The mean ratio over all orthologues for each lineage was applied as a relative rate correction to the K_s values for their respective taxa. To assess if duplications occurred after lineage divergence, one-sided Wilcoxon Rank Sum tests were used to compare the distribution of rate-corrected duplications versus the distribution of rate-corrected lineage divergences from each orthologue phylogeny.

Ages for ancient genome duplications were estimated from the mean synonymous divergence of *Camellia* and *Actinidia.* Using the maximum likelihood age estimate for the divergence time of *Camellia* and *Actinidia* from Wikström *et al.* (2001) – 71 MYA – and the mean number of synonymous substitutions among the nuclear orthologues found above, the synonymous substitution rate per million years was calculated as in Gaut and Doebley (1997). The age of each Ericales ancient genome duplication was then estimated from the data of *Actinidia deliciosa*, the species with the most complete data set, using this calibrated substitution rate based on the median peak K_s from the mixture model analyses.

Gene retention pattern analyses

GO annotations of the *Actinidia* (*Actinidia chinensis*, *A. deliciosa*, *A. eriantha*, *A. arguta*, *A. hemsleyana*, *A. polygama* and *A. setosa*) transcriptome from bud, fruit, leaf, petal, root and stem were obtained through discontiguous MegaBlast searches against *Arabidopsis thaliana* cds from TAIR (Initiative TAG, 2000) for the best hit with at least 100 bp aligned and an *e* value of $1e-10$. To ensure comprehensive coverage of the *Actinidia* transcriptome, all EST reads for each *Actinidia* species were pooled into a single assembly with TGICL, analysed with the above duplication pipeline, and annotated with the TAIR cds. To identify significant differences among GO annotations, chi-square tests with *P* values computed from 100 000 Monte Carlo simulations were conducted in R (R Development Core Team, 2005). When chi-square tests were significant ($P < 0{\cdot}05$), GO categories with residuals $>|2|$ were implicated as major contributors to the significant chi-square statistic. This statistical framework was used to evaluate differences in GO category representation patterns between paleologues (paralogues derived from paleopolyploidy) and non-paleologues (paralogues not derived from paleopolyploidy) in *Actinidia*. Expander (Shamir *et al.*, 2005) was employed to cluster different GO category patterns of normalized number of paleologues and non-paleologues using the Complete Linkage Clustering (default options). Boundaries for each whole genome duplication were defined by the mixture model results. Duplications in the region of overlap between two distributions were assigned to a particular whole genome duplication based on their probability assignment from the mixture model analysis.

RESULTS

Age distributions of gene duplications

A total of 2916 gene duplications younger than $K_s = 2$ were inferred across the total data set of 49 715 assembled unigenes (Table S1 in Supplementary data, available online). The histograms of duplication ages for each Ericales species analysed demonstrated evidence of at least one large-scale duplication (Fig. 1). Consistent with this observation, the K-S goodness-of-fit test rejected the null model of no large duplications for each taxon examined ($P = 0$). Subsequent mixture model analyses identified multiple peaks in the analysed Ericales taxa (Fig. 1, Table 1 and Table S1). In the genus *Actinidia*, the duplication distributions of *A. chinensis*, *A. deliciosa* and *A. eriantha* each contained evidence of three peaks of similar synonymous divergences. For example, in *A. chinensis* these peaks are located at median K_s of 0·13777, 0·4221 and 1·1923 (Fig. 1 and Table S1). The ΔBIC values indicate that models including these peaks describe the age distributions significantly better than models that lack these duplications (Tables 2 and S1). Similarly, evidence was found of two peaks in *Camellia* with K_s medians at 0·379355 and 1·038 (Tables 1 and S1). Although the mixture model for *Diospyros* does not identify significant peaks – probably because there are relatively few data – two peaks are apparent in the histogram near $K_s = 0{\cdot}7$ and $K_s = 1{\cdot}3$ (Fig. 1). The mixture model also identified one or two components in the duplication-rich initial peaks ($<0{\cdot}1$) of all taxa that are likely to reflect variation in birth and death rates of tandem, small-scale, segmental duplications, or alleles.

To resolve the number and phylogenetic placement of large-scale genome duplications, rate-corrected duplications were placed on the mean phylogeny of orthologues in Ericales (Fig. 2 and Tables 2 and S2). A total of 37 nuclear orthologues were identified among the analysed taxa. Mean ratios of K_s branch lengths for *Actinidia* and *Camellia* versus the *Diospyros* are 1·02 and 0·89, respectively (Table 2). Taking into account this rate heterogeneity, the mean rate-corrected divergence between *Actinidia* and *Camellia* was calculated as $K_s = 0{\cdot}4172$, whereas *Diospyros* and these two genera diverged at $K_s = 0{\cdot}5247$. After correcting the duplication peak medians with the appropriate ratio, the peaks in *Actinida*, named *Ad*-α and *Ad*-β, are centred at $K_s = 0{\cdot}16$ and $K_s = 0{\cdot}42$, respectively. Similarly, *Camellia* contains a peak centred at $K_s = 0{\cdot}42$ after rate correction. Considering the placement of these duplications on the Ericales phylogeny, the youngest peak, *Ad*-α, occurred after the divergence of *Actinidia* and *Camellia* (*U*-test $P < 1e-5$), but *Ad*-β most likely occurred prior to the divergence of these two lineages (*U*-test $P = 1$; Tables 1 and S2 and Fig. 2). The current estimates also place *Ad*-β after the divergence of *Actinidia* and *Camellia* from the clade containing *Diospyros* (*U*-test $P < 1e-5$). However, *Diospyros* may have a duplication near $K_s = 0{\cdot}75$ that could be *Ad*-β, and additional data are needed to resolve better the incidence and position of duplications in this part of the phylogeny. Further, both *Actinidia* and *Camellia* have an older peak centred at $K_s = 1{\cdot}13$ and $K_s = 1{\cdot}17$, respectively, well before divergence of the Ericales and *Populus* or *Vitis*. This peak is also apparent in the histogram of *Diospyros* at $K_s = 1{\cdot}3$. Taking into account

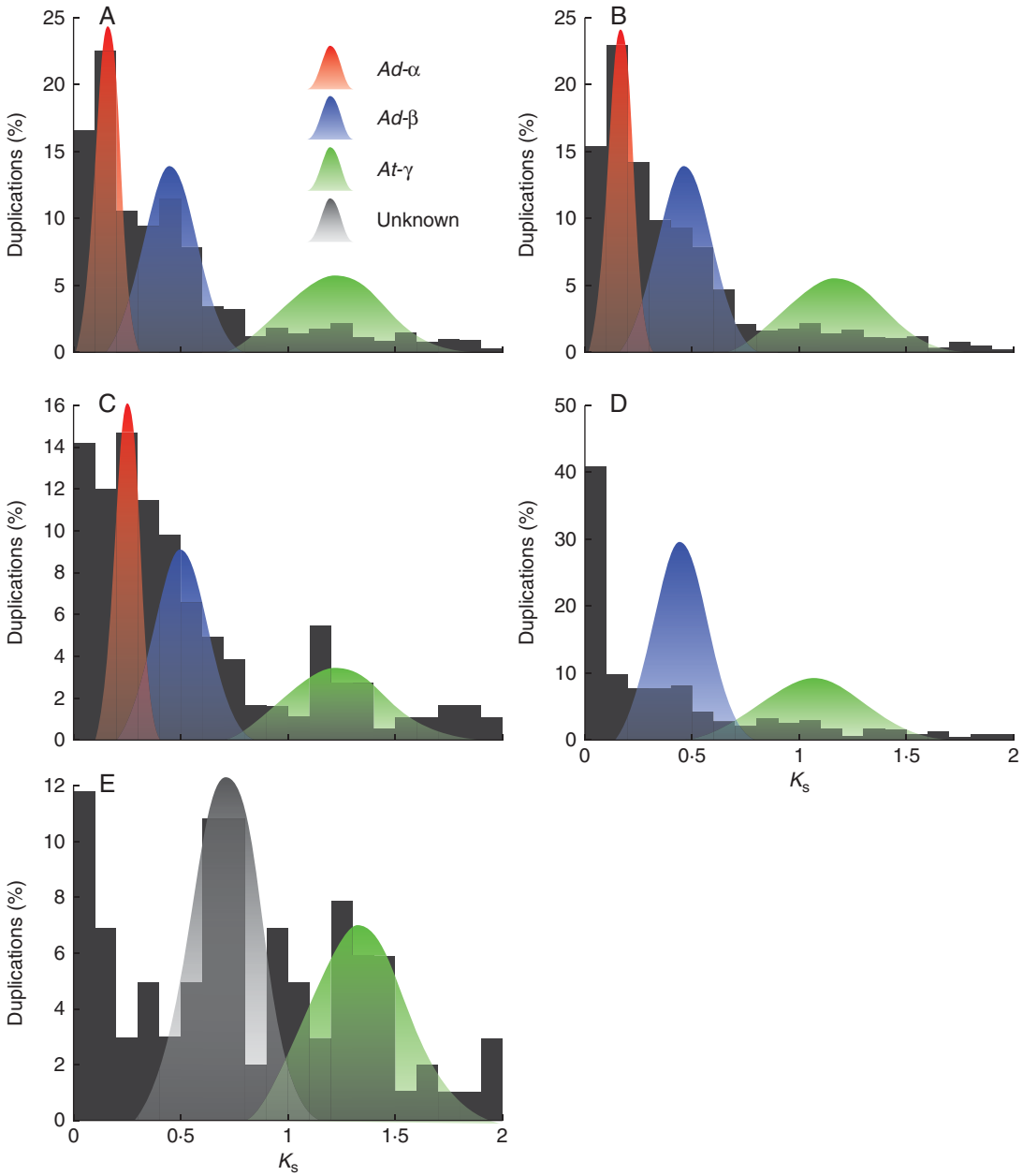

FIG. 1. Histograms of the age distribution of gene duplications from (A) *Actinidia chinensis*, (B) *Actinidia deliciosa*, (C) *Actinidia eriantha*, (D) *Camellia sinensis* and (E) *Diospyros kaki*. Shaded distributions represent mixture model fits of inferred whole genome duplications: red = *Ad*-α, blue = *Ad*-β, green = putative *At*-γ, and grey = possible duplication.

the antiquity of this age range, this peak is likely to correspond to *At*-γ, an ancient polyploidy shared by most eudicots (Vision *et al.*, 2000; Bowers *et al.*, 2003; De Bodt *et al.*, 2005; Cui *et al.*, 2006; Jaillon *et al.*, 2007; Barker *et al.*, 2008, 2009; Lyons *et al.*, 2008).

Age estimates of the ancient genome duplications are consistent with the phylogenetic placements. Based on the mean synonymous divergence across the 37 nuclear orthologue phylogenies for *Actinidia* and *Camellia* (Table S2) a synonymous substitution rate of $2{\cdot}81 \times 10^{-9}$ is estimated. Based on the peak medians for *A. deliciosa*, the species whose orthologues were used to calculate the substitution rate, the ages of the Ericales genome duplications were calculated. *Ad*-α is estimated to have occurred approx. 28·3 MYA, whereas *Ad*-β is estimated to have occurred nearly 75·9 MYA.

Comparison of gene retention and loss patterns

The GO patterns of genes retained in duplicate varied among the different large-scale duplication events. For the pooled *Actinidia* EST assembly, the GO patterns of genes retained in duplicate from the *Ad*-α, *Ad*-β and *At*-γ duplications were significantly different from each other ($\chi^2 = 497{\cdot}2$, $P = 1e - 5$; Table S3). Hierarchical clustering of normalized GO slim data

TABLE 1. *Rate-corrected mixture model medians of* Actinidia *and* Camellia *paleopolyploidizations*

Species	Relative rate (% K_s)	Rate-corrected paleopolyploidization			BIC
		Ad-α	*Ad*-β	*At*-γ	
Actinidia chinensis	102	0·135069	0·413824	1·168922	350·9
Actinidia deliciosa	102	0·155794	0·418020	1·099020	229·9
Actinidia eriantha	102	0·184069	0·443118	1·205196	169·5
Camellia sinensis	89		0·426242	1·166292	−17·8

TABLE 2. *Mixture model ΔBIC values for* Actinidia *and* Camellia *age distributions without (w/o) the inferred paleopolyploidizations*

	w/o (*Ad*-α)	w/o (*Ad*-β)	w/o (*At*-γ)
Actinidia chinensis	1183·54	Fit by all models	13·87
Actinidia deliciosa	745·40	Fit by all models	47·98
Actinidia eriantha	97·54	Fit by all models	14·41
Camellia sinensis	NA	380·68	8·58

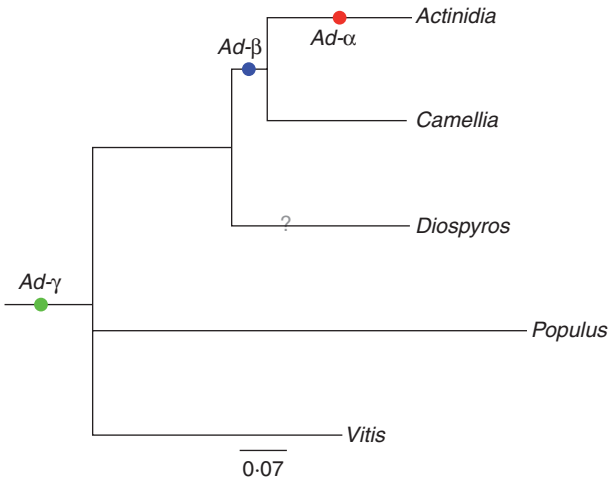

FIG. 2. Phylogeny of Ericales taxa and related Rosid outgroups displaying inferred paleopolyploidizations. Branch lengths are mean K_s values from 37 nuclear orthologues (see Table S2, available online). Coloured dots indicate inferred paleopolyploidizations placed in relation to lineage divergence base on the rate corrections. The '?' represents an ambiguous paleopolyploidization inferred from *Diospyros* ESTs.

patterns showed that non-paleopolyploid duplications and *Ad*-α duplicates were grouped together while duplicates from *Ad*-β and *At*-γ were grouped together although they were all significantly different from each other (Fig. 3). The most consistent pattern observed was enrichment of 'plastid' and 'chloroplast' GO slim categories in the non-paleopolyploid and *Ad*-α duplicates with significant under-representation of these categories in duplicates from *Ad*-β, and *At*-γ. However, compared with the non-paleologues, analyses found that the pooled paleologues from all three duplications were enriched for the 'developmental processes', 'hydrolase activity' and 'kinase activity' GO categories with 'electron transport' and 'structural molecular activity' under-represented (Fig. 3 and Table S3).

DISCUSSION

Based on cytological analyses, botanists have long suggested that many plants have experienced ancient genome duplications and genomic data are now able to test these hypotheses. In angiosperms, a long-standing estimate for the original base chromosome number ($x = 7$) is much smaller than the accepted base numbers of many genera (Pires and Hertweck, 2008). One plausible explanation for this observation is past polyploidy. A classic example of this situation occurs in the Actinidiaceae where the genus *Actinidia* has long been suspected to have a polyploid ancestry (Huang and Ferguson, 2007). Previous cytological analyses of the base numbers for the three extant genera of the Actinidiaceae – *Saurauia* ($x = 13$), *Clematoclethra* ($x = 12$) and *Actinidia* ($x = 29$) – indicated that *Actinidia* was likely to be a paleotetraploid (He *et al.*, 2005). The observation of peaks in the history of gene duplications in species of *Actinidia* is consistent with this hypothesis and provides the first genomic evidence of ancient polyploidy in the genus. The age estimate for *Ad*-α, 28·3 MYA, is consistent with this duplication being restricted to the Actinidiaceae but suggests it might be older than the genus. Additional data from the other two genera of the Actinidiaceae are needed to test further if the *Ad*-α duplication is restricted to the ancestry of *Actinidia* as expected from chromosomal analyses or is shared by more members of the family.

Recent cytological research on *Actinidia* and its relatives (reviewed in Huang and Ferguson, 2007) provides an excellent opportunity to evaluate genomic approaches for detecting ancient genome duplications. Examples such as *Actinidia* are critical because chromosomal diploidization has obscured evidence of past polyploidy in many groups. As in the Heliantheae (Barker *et al.*, 2008), the same bioinformatic tools combined with modest amounts of transcriptome data sufficiently recovered evidence of paleopolyploidy consistent with previous research. Future phylogenomic analyses of the Actinidiaceae to place *Ad*-α on the phylogeny more precisely will provide a further test of this approach. Considering the decreasing cost of transcriptome sequencing and the coming explosion of genomic data, these natural examples are critical for selecting the best combination of data and methods for inferring ancient polyploidy across the eukaryote phylogeny.

Combined with ESTs of other genera from the Ericales available on GenBank evidence was found that *Actinidia* and *Camellia* have an older shared duplication, *Ad*-β. The rate-corrected phylogeny suggests that *Ad*-β occurred after the divergence of *Diospyros* from the clade containing *Actinidia* and *Camellia.* However, a peak near this position is apparent in the duplication distribution of *Diospyros*, but it is not clear if this is a whole genome duplication or *Ad*-β because there are relatively few data points and the signal is noisy. The estimated age of *Ad*-β, 75·9 MYA, is also inconclusive because *Diospyros* diverged from the clade containing *Actinidia* and *Camellia* at nearly the same time (Wikström *et al.*, 2001). Additional transcriptome data from *Diospyros* and other Ericales are needed to confirm the location of *Ad*-β. It is worth noting that the currently proposed position of *Ad*-β is within the ancestry of a distinct clade of families

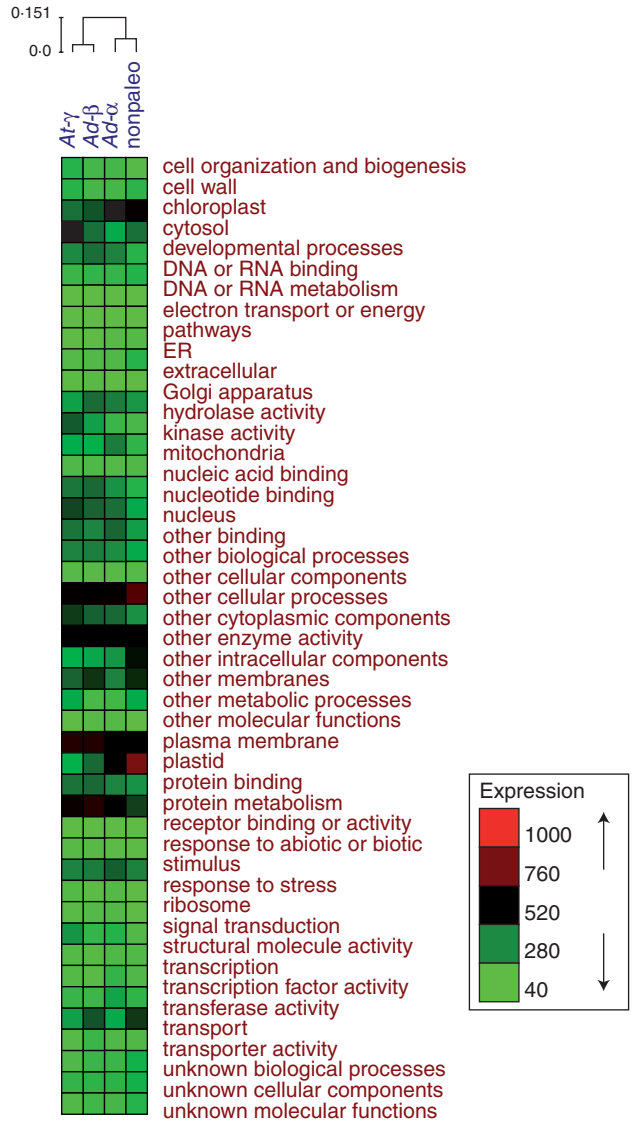

FIG. 3. GO annotations of *Actinidia* non-paleologues and paleologues. The colour enrichment represents the normalized unigene number of each GO slim category in the pooled *Actinidia* non-paleologues and paleologues. The GO slim category patterns among non-paleologues and paleologues were revealed by complete linkage hierarchical clustering. Boxes reflect relative level of GO category representation from low (green) to high (red).

that includes the Theaceae, Symplocaceae, Styracaceae, Diapensiaceae, Ericaceae and Actinidiaceae (Geuten *et al.*, 2004). A recent study of the floral regulatory genes *APETALA3* (*AP3*) and *PISTILLATA* (*PI*) genes found that the *PI* gene duplicated before the divergence of basal asteroid Ericales families but this duplication had not been shown in *AP3* lineage (Viaene *et al.*, 2009). It is possible this *PI* gene duplication is the result of *Ad*-β, but it cannot be ruled out that it was produced by random small-scale or segmental duplications. More genomic data from this clade are needed to examine which other families share the *Ad*-β duplication. These data will provide a valuable resource to test if this and possibly other paleopolyploidizations are correlated with the diversification of the Ericales as well as the association of duplications with shifts in floral regulatory gene counts and morphology.

In angiosperms, previous analyses have found the biased retention of some functional gene classes after large-scale duplication events. In particular, dosage-sensitive gene categories, such as those involved in signal transduction and transcriptional regulation, were preferentially retained after the three whole genome duplication events within the ancestor of *Arabidopsis thaliana*, whereas there was biased loss of these genes after small-scale duplication events (Blanc and Wolfe, 2004; Maere *et al.*, 2005). These patterns of duplicate gene retention and loss have been described as support for the dosage-balance hypothesis (Birchler *et al.*, 2007; Edger and Pires, 2009). According to this hypothesis, the stoichiometry of dosage-sensitive gene products must be maintained for the proper functioning of signalling networks or macromolecular complexes, especially those associated with regulatory processes (Edger and Pires, 2009). Although analyses of *Arabidopsis* paleologue support this hypothesis, results from other plant lineages have been less consistent with it. For example, the paleologues of the Compositae were over enriched for the GO categories 'structural components' and 'cellular organization' (Barker *et al.*, 2008). In the moss *Physcomitrella patens*, GO and pathway analyses of the duplicated genes reveal different biases of gene retention compared with seed plants, and enriched GO categories all belong to the KEGG ontology (KO) class 'metabolism' (Rensing *et al.*, 2007).

In the current study of *Actinidia*, no broad pattern of paleologue retention emerges from analyses of each duplication. Only non-paleologues and *Ad*-α shared some similarity of GO slim pattern in which the 'chloroplast' and 'plastid' categories are enriched whereas in *Ad*-β and *At*-γ these were reduced. However, the GO categories of non-paleologues and *Ad*-α may be confounded by the difficulty of separating genes from these two distributions because of the young age of *Ad*-α. A broad pattern of duplicate retention may also not emerge if the *Actinidia* transcriptome was not sequenced to sufficient depth to reveal a pattern. However, similar numbers of pooled transcriptome reads did reveal a consistent pattern among the Compositae (Barker *et al.*, 2008). An alternative explanation is that independent loss of paralogues derived from each duplication event in *Actinidia* obscures an overall pattern and it may be more appropriate to consider all paleologues. In this case, when compared with non-paleologues, all the paleologues (pooled from the three WGDs) have GO categories such as 'developmental processes', 'hydrolase activity', 'kinase activity' over-represented. This result is much more consistent with previous analyses from *Arabidopsis* (Blanc and Wolfe, 2004; Maere *et al.*, 2005) and the predictions of the dosage-balance hypothesis (Birchler *et al.*, 2007; Edger and Pires, 2009). This result also suggests that previous analyses which did not support the dosage-balance hypothesis for paleologue retention, particularly Barker *et al.* (2008), should be re-evaluated in this manner. However, the significant GO category consistency of paleologues retained across the multiple duplications in the Compositae suggests that the result would not be drastically different (Barker *et al.*, 2008). Regardless, future analyses of paleologue retention should examine the overall

paleologues as well as the genes retained from individual duplications to provide a more complete picture of duplicate retention biases.

Future research on the Ericales, and *Actinidia* in particular, provides an outstanding opportunity to understand better the consequences of ancient genome duplication. Additional genomic from other Ericales will permit more precise placement of the ancient duplication events and a critical evaluation of whether *Ad*-β is associated with the K-T boundary as suggested for other duplications by Fawcett *et al.* (2009). Deeper transcriptome sequencing is needed in *Actinidia* to make phylogenetic and functional genomic comparisons of each gene family in the enriched GO categories among species of *Actinidia* and to reveal the fates of those duplicated genes in groups of Ericales which share the ancient duplications. Considering that ancient genome duplications account for approx. 75 % of duplicate genes in *Actinidia* and there is variation in ploidy among species and cultivars, the genus provides a unique opportunity to understand how paleologues may uniquely contribute to domestication, the evolution of morphological complexity, and the diversification of this unique group of Ericales.

SUPPLEMENTARY DATA

Supplementary data are available online at www.aob.oxfordjournals.org and consist of the following tables. Tables S1: Gene family size descriptions and mixture model distributions for Ericales EST data sets. Table S2: Phylogenies of nuclear orthologues for Ericales with *Populus* and *Vitis* outgroup sequences. Table S3: Normalized frequency of duplicate pairs.

ACKNOWLEDGEMENTS

We thank The Horticultural and Food Research Institute of New Zealand and Max-Planck-Institute of Molecular Plant Physiology of Germany for providing the ESTs to the public. We also thank Professor Ying Wang from Wuhan Botanical Garden of CAS for productive discussions. M.S.B. is supported by the Natural Sciences and Engineering Research Council of Canada CREATE Training Program in Biodiversity Research and a Young International Scientist Fellowship from the Chinese Academy of Science (2009Y2BS3). T.S. and H.W.H. are supported by a key initiative grant of the Chinese Academy of Sciences (KSCX2-YW-N-061) and National Science Foundation of China grant (30771479) and the International Partnership Program for Creative Research Teams jointly funded by Chinese Academy of Sciences and State Administration of Foreign Experts Affairs. We also acknowledge Key Laboratory of Plant Resource Conservation and Sustainable Utilization, Chinese Academy of Sciences.

LITERATURE CITED

Altschul SF, Madden TL, Schäffer AA, *et al*. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**: 3389–3402.

Anderberg AA, Rydin C, Källersjö M. 2002. Phylogenetic relationships in the order Ericales s.l.: analyses of molecular data from five genes from the plastid and mitochondrial genomes. *American Journal of Botany* **89**: 677–687.

Barker MS, Wolf PG. 2010. Unfurling fern biology in the genomics age. *BioScience* **60**: 177–185.

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

Barker MS, Kane NC, Matvienko M, *et al*. 2008. Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Molecular Biology and Evolution* **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 Research* **24**: 2730–2739.

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

Birchler JA, Yao H, Chudalayandi S. 2007. Biological consequences of dosage dependent gene regulatory systems. *Biochemica et Biophysica Acta* **1769**: 422–428.

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

Crowhurst RN, Gleave AP, Macrae EA, *et al*. 2008. Analysis of expressed sequence tags from *Actinidia*: applications of a cross species EST database for gene discovery in the areas of flavor, health, color and ripening. *BMC Genomics* **9**: 351+.

Cui L, Wall KP, Leebens-Mack JH, *et al*. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Research* **16**: 738–749.

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

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

Edger PP, Pires JC. 2009. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Research* **17**: 699–717.

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. *Proceedings of the National Academy of Sciences of the USA* **106**: 5737–5742.

Gaut BS, Doebley JF. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. *Proceedings of the National Academy of Sciences of the USA* **94**: 6809–6814.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* **11**: 725–736.

Geuten K, Smets E, Schols P, *et al*. 2004. Conflicting phylogenies of balsaminoid families and the polytomy in Ericales: combining data in a Bayesian framework. *Molecular Phylogenetics and Evolution* **31**: 711–729.

He ZC, Li JQ, Cai Q, Wang Q. 2005. The cytology of *Actinidia*, *Saurauia* and *Clematoclethra* (Actinidiaceae). *Botanical Journal of the Linnean Society* **147**: 369–374.

Huang HW, Ferguson AR. 2007. Genetic resources of kiwifruit: domestication and breeding. *Horticultural Reviews* **33**: 1–121.

Initiative TAG. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796–815.

Jaillon O, Aury JM, Noel B, *et al*. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**: 463–467.

Lynch M, Connery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151–1155.

Lyons E, Pedersen B, Kane J, *et al*. 2008. Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with Rosids. *Plant Physiology* **148**: 1772–1781.

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

Ma B, Tromp J, Li M. 2002. PatternHunter: faster and more sensitive homology search *Bioinformatics* **18**: 440–445.

Maere S, Bodt SD, Raes J, *et al*. 2005. Modeling gene and genome duplications in eukaryotes. *Proceedings of the National Academy of Sciences of the USA* **102**: 5454–5459.

Pfeil B, Schlueter J, Shoemaker R, Doyle J. 2005. Placing paleopolyploidy in relation to taxon divergence: a phylogenetic analysis in legumes using 39 gene families. *Systematic Biology* **54**: 441–454.

Pires JC, Hertweck KL. 2008. A renaissance of cytogenetics: studies in polyploidy and chromosomal evolution. *Annals of the Missouri Botanical Garden* **95**: 275–281.

Quackenbush J, Liang F, Holt I, Pertea G, Upton J. 2000. The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucleic Acid Research* **28**: 141–145.

R Development Core Team R. 2005. *A language and environment for statistical computing, reference index version 2xx (2005)*. Vienna: R Foundation for Statistical Computing ISBN 3-900051-07-0. Available from: http: // wwwR-projectorg.

Rensing SA, Ick J, Fawcett JA, *et al*. 2007. An ancient genome duplication contributed to the abundance of metabolic genes in the moss *Physcomitrella patens*. *BMC Evolutionary Biology* **7**: 130. doi:10.1186/1471-2148-7-130

Schlueter JA, Dixon P, Granger C, *et al*. 2004. Shoemaker RC 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. *The Plant Cell* **18**: 1152–1165.

Shamir R, Maron-Katz A, Tanay A, *et al*. 2005. EXPANDER: an integrative program suite for microarray data analysis. *BMC Bioinformatics* **6**: 232. doi:10.1186/1471-2105-6-232

Soltis DE, Albert VA, Leebens-Mack J, *et al*. 2009. Polyploidy and angiosperm diversification. *American Journal of Botany* **96**: 336–348.

Sterck L, Rombauts S, Jansson S, Sterky F, Rouze P, Peer YVD. 2005. EST data suggest that poplar is an ancient polyploid. *New Phytologist* **167**: 165–170.

Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH. 2008. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Research* **18**: 1944–1954.

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

Viaene T, Vekemans D, Irish VF, *et al*. 2009. *Pistillata*: duplications as a mode for floral diversification in (basal) asterids. *Molecular Biology and Evolution* **26**: 2627–2645.

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 Acid Research* **31**: 3537–3539.

Wheeler DL, Barrett T, Benson DA, *et al*. 2007. Database resources of the National Center for Biotechnology Information. *Nucleic Acid Research* **35**: D5–D12.

Wikström N, Savolainen V, Chase MW. 2001. Evolution of the angiosperms: calibrating the family tree. *Proceedings of the Royal Society. B: Biological Sciences* **268**: 2211–2220.

Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. 2009. The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences of the USA* **106**: 13875–13879.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences: CABIOS* **13**: 555–556.

Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology* **7**: 203–214.