Available online at www.sciencedirect.com
SCIENCE DIRECT®
Molecular Phylogenetics and Evolution 33 (2004) 580–594
MOLECULAR PHYLOGENETICS AND EVOLUTION
www.elsevier.com/locate/ympev

Heterogeneous evolution of the *Myc-like* Anthocyanin regulatory gene and its phylogenetic utility in *Cornus* L. (Cornaceae)

Chuanzhu Fan[a,*], Michael D. Purugganan[b], David T. Thomas[a], Brian M. Wiegmann[c], (Jenny) Qiu-Yun Xiang[a]

[a] *Department of Botany, North Carolina State University, Raleigh, NC 27695-7612, USA*
[b] *Department of Genetics, North Carolina State University, Raleigh, NC 27695-7614, USA*
[c] *Department of Entomology, North Carolina State University, Raleigh, NC 27695-7613, USA*

Received 1 December 2003
Available online 2 October 2004

Abstract

Anthocyanin is a major pigment in vegetative and floral organs of most plants and plays an important role in plant evolution. The anthocyanin regulatory genes are responsible for regulating transcription of genes in the anthocyanin synthetic pathway. To assess evolutionary significance of sequence variation and evaluate the phylogenetic utility of an anthocyanin regulatory gene, we compared nucleotide sequences of the *myc*-like anthocyanin regulatory gene in the genus of dogwoods (*Cornus*: Cornaceae). Phylogenetic analyses demonstrate that the *myc*-like anthocyanin regulatory gene has potential as an informative phylogenetic marker at different taxonomic levels, depending on the data set considered (DNA or protein sequences) and regions applied (exons or introns). Pairwise nonsynonymous and synonymous substitution rate tests and codon-based substitution models were applied to characterize variation and to identify sites under diversifying selection. Mosaic evolution and heterogeneous rates among different domains and sites were detected.

Keywords: *Cornus*; Heterogeneous evolution; *myc*-like anthocyanin regulatory gene; Phylogenetics; Positive selection

1. Introduction

Plant species display remarkable diversity in the pattern and intensity of red or purple pigmentation. It is well understood that anthocyanins are largely responsible for the purple-red pigmentation of vegetative and floral organs in most plant species (Mol et al., 1998). Studies have indicated that these plant pigments assist in pollinator attraction, fruit dispersal, pollen viability, plant disease defense, and UV protection (Durbin et al., 2000; Epperson and Clegg, 1987; Ludwig and Wessler, 1990; Stapleton, 1992). Mutations that stop anthocyanin production are variable and often have readily observable phenotypes. Two classes of genes affect the biosynthesis of anthocyanin. One class encodes enzymes required for pigment biosynthesis (Durbin et al., 2000; Spelt et al., 2000) and the other class regulates the anthocyanin biosynthetic genes, a gene family present in diverse plant species (Goodrich et al., 1992). Previous studies have shown that the structural genes that encode the enzymes of the anthocyanin pathway are conserved among different plant species (Holton and Cornish, 1995; Quattrocchio et al., 1998). This suggests that changes in regulation that affect expression of these structural genes are at least partly responsible for the variability of pigmentation patterns observed in plants.

* Corresponding author. Present address: Department of Ecology and Evolution, The University of Chicago, 1101 E. 57th St., Chicago, IL 60637. Fax: +1 773 702 9740.
E-mail address: cfan@uchicago.edu (C. Fan).

1055-7903/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2004.08.002

Genetic studies of mutations in *Zea mays* identified two families of regulatory genes that control the transcription of all anthocyanin biosynthetic structural genes, the *R* and *c1* families. *R* family genes encode *myc*-like proteins, which contain a basic helix–loop–helix (bHLH) motif found in other eukaryotic transcriptional factors (e.g., Davis et al., 1987; DePinho et al., 1987). In maize, the *R* family includes *r*, *lc*, *sn*, and *b* genes. The *c1* family genes (e.g., *c1* and *pl*) encode *myb*-type transcription activators. Homologues of the *R* family have been identified in several dicot species including *Delila* genes in *Antirrhinum majus* (Goodrich et al., 1992), *myc-rp* and *myc-gp* in *Perilla* (Gong et al., 1999), *bHLH* transcription factor in *Arabidopsis* (Bate and Rothstein, 1997), *jaf13* in *Petunia* (Quattrocchio et al., 1998), and *ghdel65* in *Gossypium* (Matz and Burr, unpublished).

Various studies indicate that the expression of *R* genes and their homologues in maize, tobacco, *Arabidopsis*, *Petunia*, snapdragon, cotton, and tomato activate the structural genes and induce pigmentation in a wide variety of tissues, especially in flowers, and that mutations of *R*-family genes result in partial expression of anthocyanin pigments in petals (Consonni et al., 1992, 1993; Goldsbourough et al., 1996; Martin et al., 1991; Lloyd et al., 1992; Quattrocchio et al., 1993, 1998). The molecular evolution of the *R* gene family in seven grass species was examined and compared with that in dicots (Purugganan and Wessler, 1994). Four conserved functional domains were identified across the monocots and dicots, including interaction (I), acidic (A), basic helix–loop–helix (bHLH), and C-terminal (C) domains. More than one-half of the protein sequences, however, have diverged rapidly among the seven species representing diverse lineages of the grass family. Nucleotide substitutions and small insertion/deletions contribute to the diversification of variable regions (Purugganan and Wessler, 1994). Moreover, multiple copies of *R* homologues have been isolated from *Sorghum* (two copies) and *Pennisetum* (four copies). Phylogenetic analyses indicate that these multiple copies arose from gene duplication events and experienced significant sequence divergence (Purugganan and Wessler, 1994).

It is generally held that regulatory genes evolve faster than structural protein genes (e.g., Purugganan, 1998; Purugganan and Wessler, 1994; Ting et al., 1998). For example, comparison of pairwise distances between duplicated structural and regulatory genes in the maize genome indicates that the ratios of nonsynonymous versus synonymous substitutions in regulatory genes are much higher than those of structural genes (see review by Purugganan, 1998). This is also well demonstrated in some Hawaiian species (e.g., Hawaiian silversword alliance, Barrier et al., 2001), suggesting that the accelerated evolution rates in floral regulatory genes may be correlated with the rapid morphological diversification in plants. However, it still remains unclear as to how rapid molecular evolution in regulatory genes might affect morphological diversification. In the present study, we examine the pattern and rate of evolution of the anthocyanin regulatory gene in a dicot group to better understand its evolution across the four identified functional domains.

Understanding the evolution of regulatory genes also provides a foundation for their use in phylogenetic analysis. Although chloroplast DNA and nuclear ribosomal DNA markers have been used extensively to generate phylogenetic hypotheses in plants, additional DNA markers from the nuclear genome are needed to resolve phylogenies at lower taxonomic levels. Recent studies indicate that single- or low-copy nuclear genes in plants are a rich source of phylogenetic information at different levels. This includes 'lower' taxonomic levels, such as interspecific relationships and the origin of allopolyploids in plants (Sang, 2002). If regulatory genes do indeed evolve more rapidly than structural genes (see above), it is likely that these genes may be good candidates for investigating phylogenetic relationships in plants.

Here, we compare homologues of the *myc*-like anthocyanin regulatory gene among dicots and monocots and among subgroups and species in the dogwood genus *Cornus* to assess the phylogenetic utility of the gene at different taxonomic levels. The dogwoods display various flower colors, including white, yellow, and purple. Using the dogwoods as an example may provide insights into the relationship between phenotypic diversification and molecular evolution of the anthocyanin regulatory genes. Moreover, dogwoods (*Cornus*) have been examined for two chloroplast genes and one nuclear ribosomal DNA gene in previous phylogenetic analyses (Fan and Xiang, 2001, 2003; Xiang et al., 1998). Molecular data from these two chloroplast genes (*matK* and *rbcL*) and one nuclear gene (26S rDNA) are available for the dogwoods from previous studies (Fan and Xiang, 2001; Xiang et al., 1998), thereby permitting comparisons among genes and between genomes. The goals of this study are: (1) to identify and characterize homologues of the *myc*-like anthocyanin regulatory gene in dogwoods; (2) to evaluate the phylogenetic utility of this gene in dogwoods; and (3) to examine the rate and pattern of sequence evolution of this regulatory gene.

2. Materials and methods

2.1. Identification and characterization of genes in Cornus

PCR primers and amplification. In order to isolate the nuclear DNA sequences of this gene in *Cornus*, degenerate primers (F1 and R2) were designed from published

sequences of the *myc*-like anthocyanin regulatory gene sequences of dicots (forward primer F1-CAATGGAGY TATRTYTTHTGGTC and reverse primer R2-TCRG TRAGRTCTTCWGGDGATAATGC). The primer F1 is located at the beginning of the 5'-end of the interaction domain (exon 1), and R2 is at the middle of the interaction domain (exon 2). The relative positions of these two primers are marked in Fig. 1. These primers were used for initial PCR and sequencing of the dwarf dogwoods. The PCR reaction using these two primers generated an 800-bp length fragment. Cloning of the PCR products revealed two types of sequences (designated as Type 'A' and Type 'B'); both are highly similar to the anthocyanin regulatory gene in *Petunia*, *Perilla*, and *A. majus* based on Blast searches at GenBank. Type-specific primers were subsequently designed to amplify and sequence a single type (type A). The type 'A' sequence was elongated via sequential PCR reactions using sequential locus specific forward primers and locus specific/degenerate reverse primers. Sequences of flanking regions and two ends of the gene, which could not be obtained via standard PCR were obtained using thermal asymmetric interlaced (TAIL) PCR (Liu and Whittier, 1995). The entire nucleotide sequence of this gene was obtained for three species of dwarf dogwoods and *C. florida* using the methods described above. "Universal" locus-specific primers within *Cornus* were then designed based on the dwarf dogwoods and *C. florida*. In the present study, we compare the sequences of Type "A" for nine species of *Cornus* representing several different subgenera.

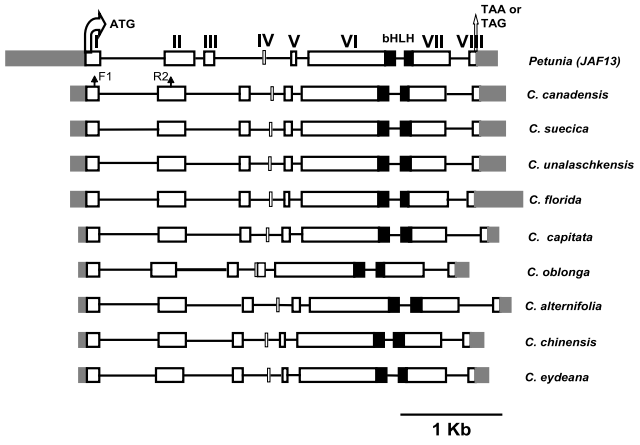

Fig. 1. Schematic map showing the overall structure of the *myc*-like anthocyanin regulatory gene (*R* homologue) for *Cornus* and *Petunia hybrida* (*JAF13*) as deduced from the full-length nucleotide sequences. The position of the ATG translation start and the TAA or TAG translation stop codons are indicated. Boxes represent the exons, and line stands for introns. Eight exons are ordered using Roman numerals (I-VIII). Seven insertion-deletions among *Cornus* species were found at exon VI through exon VIII (see Supplementary Materials for details). The region encoding bHLH is shown in dark, which is located in exon VI and VII. The flanking regions are shown as shaded. Two primers (F1 and R2) for initial PCR are marked. The genomic sequences for *Cornus* are available through the GenBank database (Accession Nos. AY465415–AY465425) and the genomic sequence of JAF 13 was kindly provided by Francesca Quattrocchio (Department of Genetics, Vrije Universiteit, De Boelelaan 1087, 1081 HV Amsterdam, The Netherlands).

Using the "universal primers," the entire sequences of the *myc*-like anthocyanin regulatory gene can be amplified for dwarf dogwoods (*C. canadensis*, *C. suecica*, and *C. unalaschkensis*), *C. florida*, and *C. capitata* using primer combinations of F0A-R2A2, F2A (or F2A1)-R3', F4A-R4A, F6A-R7A, and F7A2-R9A. PCR amplification for four other species (*C. oblonga*, *C. eydeana*, *C. alternifolia*, and *C. chinensis*) was achieved using four other combinations of primers [F0A2-R2A3 (or R2), F1-R3A, F3'-R8A3, and F7A2-R8A2]. For all primer combinations, the adjacent amplified fragments overlap by at least 50 bp at two ends. A gel extraction procedure (1.5% agarose gel electrophoresis followed by purification using a QIAquick PCR purification kit from Qiagen, Maryland 20874, USA) or TOPO TA cloning was applied for some cases where multiple PCR bands were obtained. Both strands of DNA were sequenced. Detailed information for locus-specific primers used for PCR amplification and sequencing is listed in Supplementary Materials. All degenerate primers and locus-specific primers were designed as described above and synthesized by IDT (Integrated DNA Technologies, INC. 1710 Commercial Park, Coralville, IA 52241-9802, USA) or Sigma Genosys (1442 Lake Front Circle, The Woodlands, TX 77380-3600, USA).

Total DNAs (as PCR templates) were extracted from fresh or silica-gel dried leaves. The protocol of DNA extraction was described previously by Xiang et al. (1998). PCR reactions were performed using different combinations of the forward and reverse primers described above. PCR reactions contained the following: 5 μL of 10× Mg^{2+} free buffer, 6 μL of 25 mmol/L $MgCl_2$, 6–10 μL of 2.5 mmol/L dNTPs, 0.5 μL of 20 μmol/L forward primer, 0.5 μL of 20 μmol/L reverse primer, 5 μL of DMSO (dimethyl sulfoxide), 1–5 μL of BSA (Bovine serum albumin, 10 mg/ml), 0.3 μL of *Taq* polymerase (Promega), 5–10 μL of 20 ng/μL total DNA extract, and calibrated to a final volume of 50 μL using sterile deionized water. In order to avoid non-specific primer annealing and to increase yield of PCR products, a hot-start (6 min of 96 °C incubation) was processed before adding *Taq* polymerase. The PCR reaction mix was run on a PTC-100 thermal cycler (MJ Research Inc., Watertown, MA, USA) as follows: (1) 94 °C for 30 s for one cycle; (2) 30–40 cycles of 94 °C for 45 s, 50–60 °C (annealing temperature optimized based on the Tm of primers) for 1 min, 72 °C for 1.5–2.5 min; (3) a terminal phase at 72 °C for 6 min. TAIL PCR was conducted using primers specific to species of dwarf dogwoods and *C. florida* with high annealing temperature and three previously published arbitrary degenerate (AD) primers (see Supplementary Materials). The three arbitrary degenerate

primers (AD1, AD2, and AD3) and the procedures of the three consecutive PCR reactions for TAIL-PCR were described previously (Liu and Whittier, 1995; Liu et al., 1995).

TOPO TA cloning. For PCR products that did not give clean sequences, TOPO TA cloning was used to isolate the different types of sequences. The PCR products were purified and cloned in competent *E. coli* cells using TOPO TA cloning techniques (Invitrogen Life Technologies, Carlsbad, CA 92008, USA). The growing colonies were screened for positive transformants using PCR amplification by T3 and T7 primers located within the vector. Ten to twenty positive transformants were inoculated to multiply the cells. Plasmid DNAs were extracted and purified using Promega Minipreps DNA purification system (Promega, Madison, WI 53711-5399, USA). The purified plasmid DNA products were directly sequenced.

Sequencing. The double-stranded (DS) PCR products were cleaned using 20% PEG (polyethylene glycol) 8000/2.5 mol/L NaCl (Morgan and Soltis, 1993; Soltis and Soltis, 1997). Purified PCR or plasmid products were used as the templates for sequencing using the ABI PRISM dRhodamine Terminator Cycle Sequencing Ready Reaction Kit (Applied Biosystems, Foster City, CA, USA). Cycle-sequencing reactions (10 μL) were prepared by combining 2 μL terminator ready reaction mix, 2 μL sequencing buffer (200 mmol/L Tris–pH 8.0, 5 mmol/L $MgCl_2$), 0.6 μL primer (5 μmol/L), 2 μL or 4 μL of 200 ng/μL cleaned PCR product or plasmid DNA, 0.5 μL DMSO, and 2.9 μL (for PCR product reactions) or 0.9 μL (for plasmid product reactions) DI water. Cycle-sequencing was conducted on a PTC-100 Programmable Thermal Controller (MJ Research Inc., Watertown, MA, USA) as follows: 25 cycles of 96 °C for 30 s, 50 °C for 15 s, and 60 °C for 4 min. Products of cycle-sequencing were cleaned using ethanol/sodium acetate precipitation (ABI applied Biosystems, Foster City, CA 94404, USA) with an additional 95% ethanol wash. The cleaned sequencing products were analyzed on an ABI-377 automated sequencer (Applied Biosystems, Foster City, CA 94404, USA). The sequence chromatogram output files for all samples were checked and edited base by base manually before being aligned.

2.2. Assessing phylogenetic utility

Species sampling. Eleven DNA samples from nine species of *Cornus* representing major clades and different subgenera of the genus were analyzed (Table 1). These included three species of dwarf dogwoods (*Cornus* subgen. *Arctocrania*: *C. canadensis*, *C. suecica*, and *C. unalaschkensis*), *C. florida*, *C. capitata*, *C. oblonga*, *C. alternifolia*, *C. eydeana*, and *C. chinensis*. These nine species represent the four major lineages of *Cornus*, the dwarf dogwoods, big-bracted dogwoods, cornelian cherries, and blue- or white-fruited dogwoods (Eyde, 1988; Fan and Xiang, 2001; Xiang et al., 1993, 1998).

Sequence alignment and phylogenetic analyses. Both DNA and protein sequences were aligned using Clustal X (Thompson et al., 1997), and adjusted manually. The amino acid sequences for species of *Cornus* were translated from DNA sequences using DNA Strider1.1 (Marck, 1988). The protein sequences of homologues of the *myc*-like/*R* anthocyanin regulatory genes for *Arabidopsis*, *Petunia*, *Gossypium hirsutum*, *Perilla*, *A. majus*, *Z. mays*, and *Oryza* downloaded from GenBank were aligned with those from *Cornus*. Phylogenetic analyses were performed using a broad protein data set including sequences of the other dicots and monocots from GenBank, and a narrow DNA data set for *Cornus* only. Only DNA sequences from the coding region were used in the phylogenetic analyses due to ambiguity of alignment in the intron regions. Both parsimony and maximum likelihood (ML) methods were used. Parsimony analyses for both protein and DNA sequence matrices were performed using PAUP* 4.0b10 (Swofford, 2002). For parsimony analysis, gaps were coded as missing

Table 1
Sampling information and GenBank accession numbers

Subgroup	Subgenus	Species	Voucher and collection locality	GenBank Accession Nos.
Dwarf dogwood	*Arctocrania*	*C. canadensis*	6-1, Xiang and Fan, 2000, British Columbia, Canada	AY465415
	Arctocrania	*C. suecica*	43-2, Xiang and Fan, 2000, Alaska, USA	AY465417
	Arctocrania	*C. suecica*	94-388, Chris Brochmann, Norway.	AY465418
	Arctocrania	*C. unalaschkensis*	2-6, Xiang and Fan, 2000, Idaho, USA	AY465416
Big-bracted dogwood	*Cynoxylon*	*C. florida*	02-16, Xiang, 2002, Veracruz, Mexico.	AY465419
	Cynoxylon	*C. florida*	02-36, Fan, 2002, North Carolina, USA	AY465420
	Syncarpea	*C. capitata*	02-188, Xiang, 2002, Weisi County, China	AY465421
Cornelian cherry	*Sinocornus*	*C. chinensis*	02-83, Xiang, 2002, Sichuan, China	AY465423
	Sinocornus	*C. eydeana*	02-232, Xiang, 2002, Yunnan, China	AY465425
Blue- or white-fruited dogwood	*Mesomora*	*C. alternifolia*	01-189, Xiang and Fan, 2001, Smoky Mountains, Tennessee, USA	AY465424
	Yinquania	*C. oblonga*	02-254, Xiang, 2002, Yunnan, China	AY465422

data, and multiple states were treated as uncertainty. Heuristic searches were performed using the MULPARS option with characters equally weighted, character states unordered, random taxon addition with 1000 replicates, and tree-bisection-reconnection (TBR) branch-swapping algorithm. ML analysis of DNA sequences incorporated the best fit model of sequence evolution estimated by Modeltest (Posada and Crandall, 1998) was conducted using PAUP* 4.0b10 and heuristic searches with 100 replicates of random taxon addition. To evaluate clade support in both parsimony and ML analyses, bootstrap analysis with 10000 replicates (Felsenstein, 1985) was performed using fast heuristic search and TBR branch-swapping.

2.3. Rate and pattern of gene evolution

Analysis of synonymous (K_s) and nonsynonymous (K_a) substitution rate. To examine the pattern and rates of nucleotide substitutions in the coding region, pairwise synonymous (K_s) and nonsynonymous (K_a) nucleotide substitution rates in the coding regions were examined. K_s and K_a were estimated using the Jukes-Cantor (Jukes and Cantor, 1969) distance model with the Nei-Gojobori method (Nei and Gojobori, 1986), implemented in *MEGA* version 2.1 (Kumar et al., 2001). The pairwise ratio of K_a/K_s was calculated by dividing the nonsynonymous nucleotide substitutions rate with the synonymous nucleotide substitutions rate. K_a and K_s were also compared for each of the four functional domains (interaction domain, acidic domain, bHLH domain, and C-terminal domain). The ratio of nonsynonymous to synonymous rate (K_a/K_s) was plotted against the rate of synonymous substitution (K_s) to examine the relationship between nonsynonymous and synonymous substitutions.

Identification of positively selected amino acid sites. Analyses of heterogeneous selection pressure at amino acid sites have been used to detect mosaic rates of evolution in protein genes. Amino acid sites in a protein under different selective pressures are indicated by their heterogeneous ω (the ratio of nonsynonymous/synonymous substitution rate, denoted as $\omega = K_a/K_s$) ratio among sites (Nielsen and Yang, 1998). We estimated the value of ω for amino acid sites using the codon-substitution model of Yang and colleagues (Yang et al., 2000). Analyses were implemented using the program Codeml of PAML (Yang, 1997). Various models of heterogeneous ω ratios among sites, including one-ratio (M0), neutral (M1), selection (M2), discrete (M3), beta (M7), and beta & ω (M8), were applied (Yang et al., 2000). Three pairs of model comparisons (M1 and M2, M0 and M3, M7 and M8) were made to determine the selection pressure of the gene. The significance of comparisons was determined by Likelihood-ratio test (LRT). The LRT test contrasts twice the log-likelihood difference with a χ^2 distribution with the degrees of freedom v equal to the difference in the number of parameters between two models (Yang et al., 2000).

3. Results

3.1. Structural characteristics

The results from Blast searches against GenBank and comparisons of amino acid alignments indicated that the protein sequences obtained in *Cornus* are highly similar to the sequences of *myc*-like/*R* anthocyanin regulatory genes in *Petunia*, *Arabidopsis*, *Perilla*, *Gossypium*, and *A. majus*. Entire nucleotide sequences of the *myc*-like anthocyanin regulatory gene in nine *Cornus* species examined were determined to be 3.5–3.75 kb. The gene contains eight exons and seven introns in most species of *Cornus* (Fig. 1). In *C. oblonga* the gene lacks intron 4 (Fig. 1). The exons are highly variable in size, ranging from only 15 bp (exon IV) to over 800 bp (exon VI). However, the sizes of exons are mostly highly conserved

Table 2
Insertion-deletion of the *myc*-like anthocyanin regulatory gene (coding region) identified in nine *Cornus* species

Indel	Exon location	Domain	Position in sequence	Species with sequences	Length (bp)	Sequence (5′–3′)
1	VI	Acidic	640–642	*C. unalaschkensis*	3	TAT
2	VI	Acidic	715–726	*C. suecica*, *C. unalaschkensis*, *C. florida*, *C. capitata*, *C. oblonga*, *C. alternifolia*, *C. chinensis*, *C. eydeana*	12	CTTGATSYDGMY
3	VI	Acidic	877–879	*C. oblonga*, *C. alternifolia*	3	ATT
4	VI	Acidic	958–969	*C. oblonga*, *C. alternifolia*	12	ATTGGTGGCTCT
5	VII	bHLH/C-terminal	1474–1491	*C. florida*, *C. capitata*, *C. oblonga*, *C. alternifolia*, *C. chinensis*, *C. eydeana*	18	TGCARGGAGSWRRCARAK
6	VII	C-terminal	1558–1560	*C. florida*, *C. capitata*	3	RAC
7	VIII	C-terminal	1906–1908	*C. canadensis*, *C. suecica*, *C. unalaschkensis*, *C. florida*, *C. capitata*, *C. alternifolia*, *C. chinensis*, *C. eydeana*	3	TGY

among species of dogwoods with no difference in exon I to exon V among the nine species. Seven indels, however, were detected from exon VI to exon VIII with four in exon VI, two in exon VII, and one in exon VIII (Table 2). The sizes of introns are highly variable in these *Cornus* species (see Supplementary Materials and Fig. 1).

Table 3
Absolute pair-wise distance matrix among nine *Cornus* species (the number of nucleotide differences)

Species	*C. canadensis*	*C. unalaschkensis*	*C. suecica*	*C. florida*	*C. capitata*	*C. oblonga*	*C. alternifolia*	*C. chinensis*
C. unalaschkensis	12							
C. suecica	43	39						
C. florida	176	179	188					
C. capitata	158	166	176	56				
C. oblonga	300	302	308	264	262			
C. alternifolia	309	312	317	273	273	73		
C. chinensis	182	187	194	161	156	253	261	
C. eydeana	204	207	214	158	153	251	266	59
Average	173.00	198.86	232.83	182.40	211.00	192.33	263.50	59
Total average	202.63							

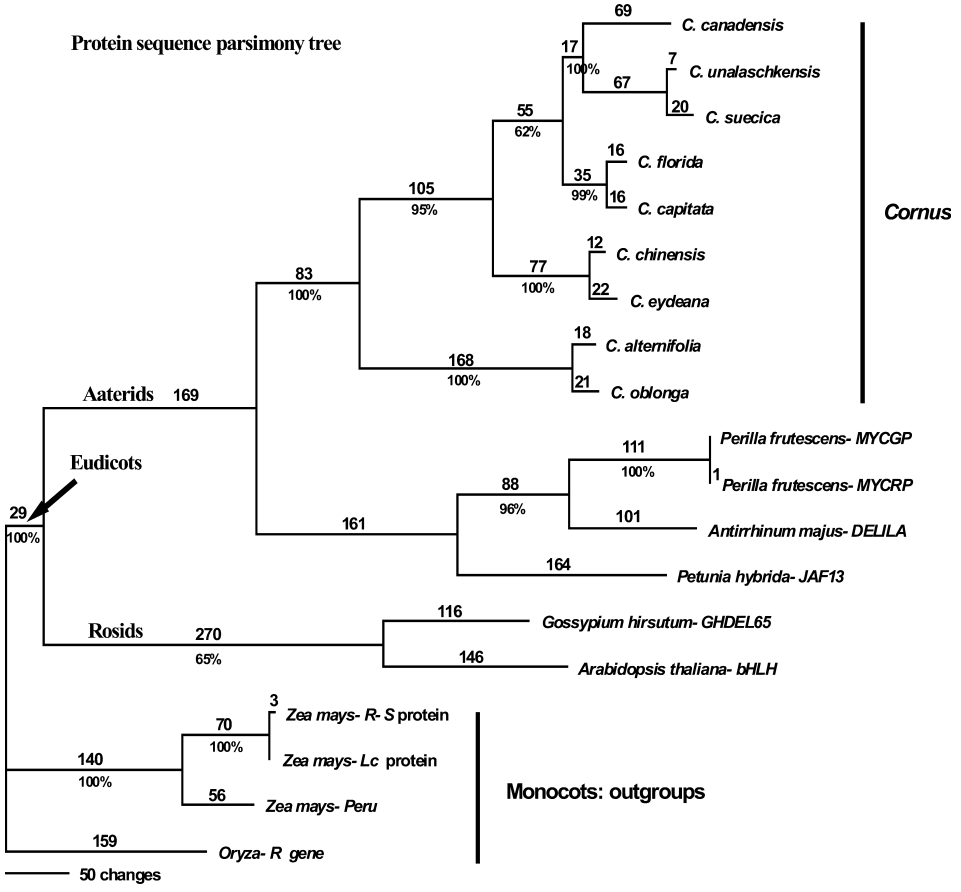

Fig. 2. One of four parsimonious trees of amino acid sequences of the *myc*-like anthocyanin regulatory gene from *Arabidopsis*, *Petunia*, *Perilla*, *Gossypium*, *Z. mays*, *Oryza*, *Antirrhinum majus*, and *Cornus*. The sequences of *Cornus* were generated in this study, and all other sequences were downloaded from GenBank. The references and GenBank accession numbers for them are listed as follows: *Arabidopsis thaliana*-bHLH (Bate and Rothstein, 1997. GenBank Accession No. AF013465); *G. hirsutum ghdel* (Matz and Burr, unpublished. GenBank Accession No. AF336280); *Petunia hybrida jaf* 13 (Quattrocchio et al., 1998. GenBank Accession No. AF020545). *Antirrhinum majus delila* (Goodrich et al., 1992. GenBank Accession No. M84913). *Perilla frutescens mycrp*, *mycgp* (Gong et al., 1999. GenBank Accession No. AB024050-Myc-rp, AB024051-Myc-gp). *Z. mays* (*Lc*, Ludwig et al., 1989. GenBank Accession No. M26227; Zm *B-Peru*, Radicella et al., 1991; GenBank Accession No. X57276; *R-S*-protein, Perrot and Cone, 1989; GenBank Accession No. X15806); *Oryza R* (Hu et al., 1996. GenBank Accession No. U39860). The sequences from maize and rice were treated as outgroups. The analysis was performed with PAUP4.0b10. Base substitutions are indicated above branches; bootstrap values are marked below branches. Tree length = 2060; CI = 0.839; RI = 0.828.

3.2. Phylogenetic utility

Sequence variation. Protein sequences were aligned among dicots including *Cornus* and five additional dicot species and two monocot species. This data matrix contains 721 sites. Among them, 578 sites (80.17%) are variable, and 497 sites (68.93%) are parsimony informative. Among only *Cornus* and other dicots, 496 (68.80%) of 721 sites are variable, and 375 (52.01%) of 721 sites are phylogenetically informative.

The DNA sequences of intron regions are highly variable within *Cornus* and can be aligned only among the closely related species (e.g., three species of dwarf dogwoods; *C. florida* and *C. capitata*). In contrast, the exon regions can be aligned easily among the nine *Cornus* species examined. The matrix of exon regions of *Cornus* contains 1908 bp and has 504 variable sites (26.42%) and 390 (20.44%) parsimony-informative sites. These values are significantly higher than those for 26S rDNA (11.56% and 4.05%, respectively) and chloroplast protein-coding genes (e.g., *matK* and *rbcL*, 9.96% variable sites and 2.77% parsimony informative sites) for *Cornus* for the same suite of taxa. The absolute pairwise distances among the nine species for exons range from 12 (between *C. canadensis* and *C. unalaschkensis*) to 317 (between *C. suecica* and *C. alternifolia*) with an average of 203 (Table 3).

Phylogenetic analyses. Phylogenetic analyses of the entire protein sequence and of just the bHLH domain both support the monophyly of *Cornus* within the asterids (Figs. 2 and 3) as is expected based on current estimates of angiosperm phylogeny (Soltis et al., 2000). The eudicots (including 14 species) are strongly supported by bootstrap analyses in both trees (100% in entire gene tree and 95% in bHLH domain tree; Figs. 2 and 3). Relationships among the major subgroups within *Cornus* based on the entire protein are consistent with results from previous studies (Fan and Xiang, 2001).

Phylogenetic analyses of nucleotides from the entire coding region were completely resolved regarding relationships among *Cornus* species. A single minimum-length tree was found in the parsimony analysis (Fig. 4). The topology of this tree is identical to that found in ML analysis. Estimates of relationships among the

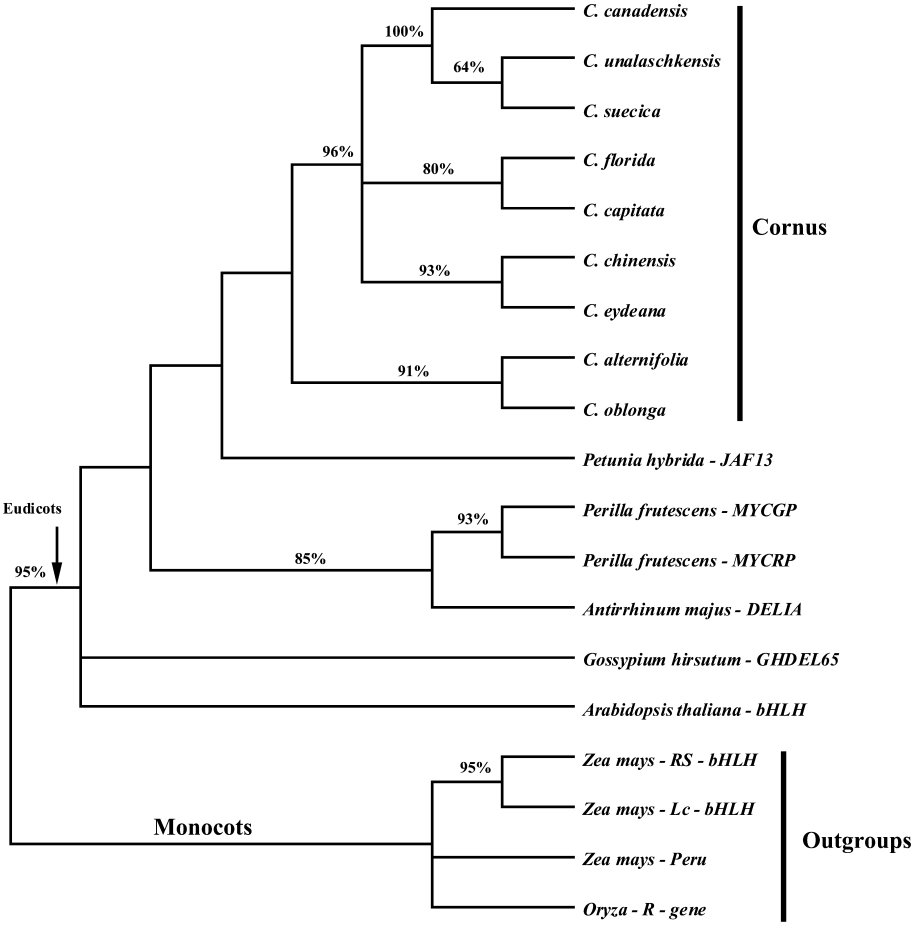

Fig. 3. Strict consensus of twelve most parsimonious trees of just bHLH domain of the *myc*-like anthocyanin regulatory gene from *Arabidopsis*, *Petunia*, *Perilla*, *Gossypium*, *Zea mays*, *Oryza*, *Antirrhinum majus*, and *Cornus* (see Fig. 2 for detailed information for genes). The analysis was performed using PAUP4.0b10. Bootstrap values are given above branches. Tree length = 122; CI = 0.820; RI = 0.848.

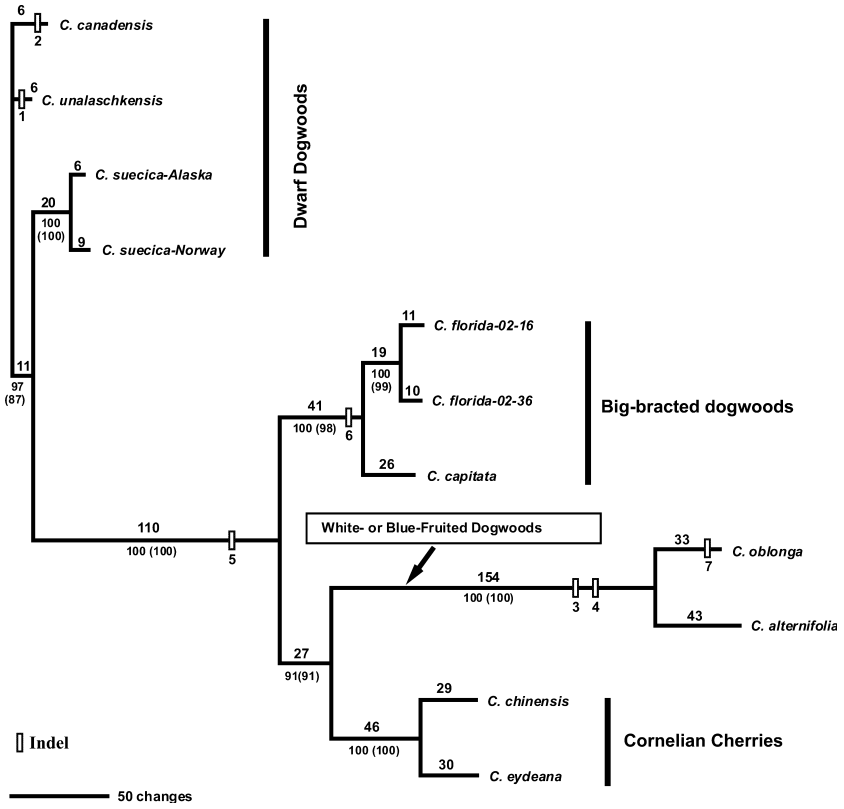

Fig. 4. The single most parsimonious unrooted tree inferred from exon sequences of eleven taxa of nine *Cornus* species (tree length = 632; CI = 0.877; RI = 0.899). Base substitutions are indicated above branches; bootstrap values for both parsimony and ML tree (parentheses) are given below branches. Seven indels identified from exons are marked (see Supplementary Materials for details).

four major lineages of *Cornus* (dwarf dogwoods, big-bracted dogwoods, cornelian cherries, and blue- or white-fruited dogwoods) were congruent with those inferred from previous 26S rDNA and combined chloroplast & 26S rDNA data analyses (Fan and Xiang, 2001, 2003). However, major dogwood lineages, and the relationships among them, are much more strongly supported by the anthocyanin regulatory gene tree than by any previous gene-based analysis (Fan and Xiang, 2001).

3.3. Pattern and rates of gene evolution

Rates of K_a and K_s. Pairwise synonymous substitution rates range from 0.05 to 0.37 with a mean value of 0.23 in *Cornus* and are higher than pairwise nonsynonymous substitution rates, which range from 0.02 to 0.15 with an average of 0.093. The mean ratio of K_a/K_s across all exons is 0.407 ± 0.040 (mean $\pm$ SD), similar to those for three of the four domains [acidic domain (0.475 ± 0.076), bHLH domain (0.383 ± 0.117), and C-terminal domain (0.423 ± 0.119)] (Fig. 5). However, the K_a/K_s ratio in the interaction domain (0.201 ± 0.095) is only about half of these values, much lower than those for the entire gene and the other three functional domains (Fig. 5). Furthermore, plots of K_a/K_s versus K_s indicate that the ratio of K_a/K_s is positively related to K_s in the acidic domain and C-terminal domain, but negatively related to K_s in the interaction and bHLH domains (Fig. 6). These data suggest that K_a increases at a higher rate than K_s in the C-terminal and acidic domains, but K_a increases at a lower rate than K_s in the interaction and bHLH domains, where there may be greater functional constraint. The individual pairwise K_a and K_s values and individual ratio of K_a/K_s are listed in Supplementary materials.

Sites under diversifying selection. The results of parameter estimation for different models using PAML with codeml are displayed in Table 4. Tests carried out using the codon-based substitution model indicate that the strictly neutral model (M1) fits the data better than the one-ratio model (M0) (Table 4). The LRT statistic for comparison of the neutral (M1) and selection models (M2) rejects M1 in favor of M2 (Table 5). However, applying the selection model (M2), we do not detect positive selection in this data set. This is probably due to the fact that the strict neutral model (M1) on which it is based is unrealistic, and the extra category added in M2 optimally accounts for deleterious mutations (with $\omega_2 = 0.17$).

The one-ratio model (M0) is rejected by a big margin when compared with model 3 (discrete model) (Table 5). This test suggests that sites under positive selection are present in this gene. Application of the discrete model (M3) suggests that a large proportion of sites ($p_2 = 43\%$, total 241 codon sites) are potentially posi-

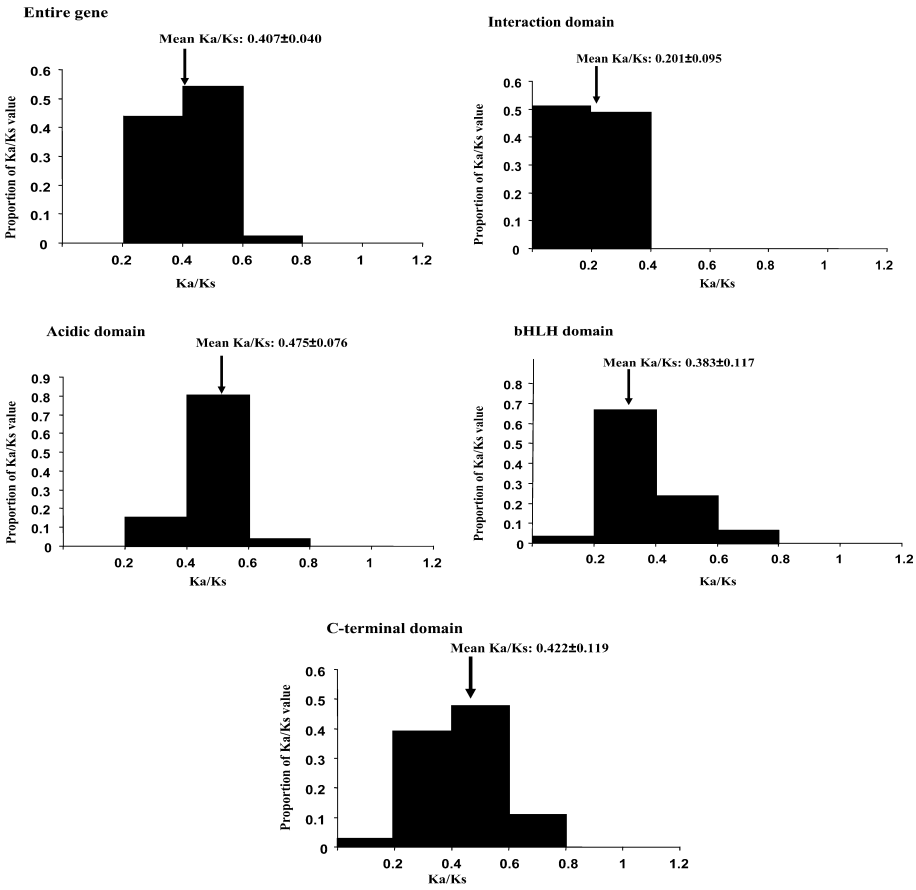

Fig. 5. Distribution of ratio of K_a/K_s for entire gene, interaction domain, acidic domain, bHLH domain, and C-terminal domain for 11 samples of nine *Cornus* species. Mean values and standard deviations are indicated. Pairwise comparisons that had synonymous substitutions below 0.05 are not shown in the histograms and were not included in the analyses.

tively selected sites with $P(\omega > 1) > 0.5$, among them 27 sites with $P(\omega > 1) > 0.95$, and 6 sites with $P(\omega > 1) > 0.99$ (Table 6). Similarly, tests with model M8 (beta & ω) suggest that 42%, or a total of 240 codon sites, are under positive diversifying selection with $P(\omega > 1) > 0.5$, among them 27 sites with $P(\omega > 1) > 0.95$, and 5 sites with $P(\omega > 1) > 0.99$ (Table 6).

Over half of 241 (240 for M8) sites with $P(\omega > 1) > 0.50$ detected are located in the acidic domain (Table 6). Seventeen of 27 sites with $P(\omega > 1) > 0.99$ are also found in the acidic domain (Table 6). Among six sites detected by M3 with $P(\omega > 1) > 0.99$, four are located in the acidic domain, and one resides in the interaction and bHLH domain, respectively (Tables 6 and 7). Site 442 in the bHLH domain is not supported with $P > 0.99$ by the M8 (Table 7).

4. Discussion

4.1. Regulatory genes in plant phylogenetics

Nuclear genes are desirable markers in plant phylogenetics at lower taxonomic levels (see review by Sang, 2002). However, only a few low- or single-copy nuclear genes have been applied to phylogenetic analyses in plant systematics (e.g., *adh*-Sang et al., 1997; Sang and Zhang, 1999; Small et al., 1998; Small and Wendel, 2000; *phyB*-Mathews and Sharrock, 1996; Mathews and Donoghue, 1999; Mathews et al., 2000; *waxy*-Mason-Gamer et al., 1998; *PgiC*-Gottlieb and Ford, 1996; *G3pdh*-Olsen and Schaal, 1999). The frequent necessity of additional procedures, however, such as extensive cloning, sequencing, and restriction endonuclease cutting, has restricted the widespread use of nuclear genes. Our study demonstrates that locus-specific primers can be designed to amplify single-copy gene sequences. Our locus-specific primers for the *myc*-like anthocyanin regulatory gene largely eliminate the need for cloning to obtain clean and unambiguous sequences. Recent studies of some nuclear regulatory genes [e.g., Floral homeotic genes (Barrier et al., 1999); two MADS-box genes: *Pistillata* (Bailey and Doyle, 1999) and *Leafy* (Nishimoto et al., 2003)] also demonstrate that regulatory genes have potential utility in plant phylogenetics and systematics. Our study adds additional evidence on regulatory genes as phylogenetic markers. The *myc*-like anthocyanin regulatory gene contributes

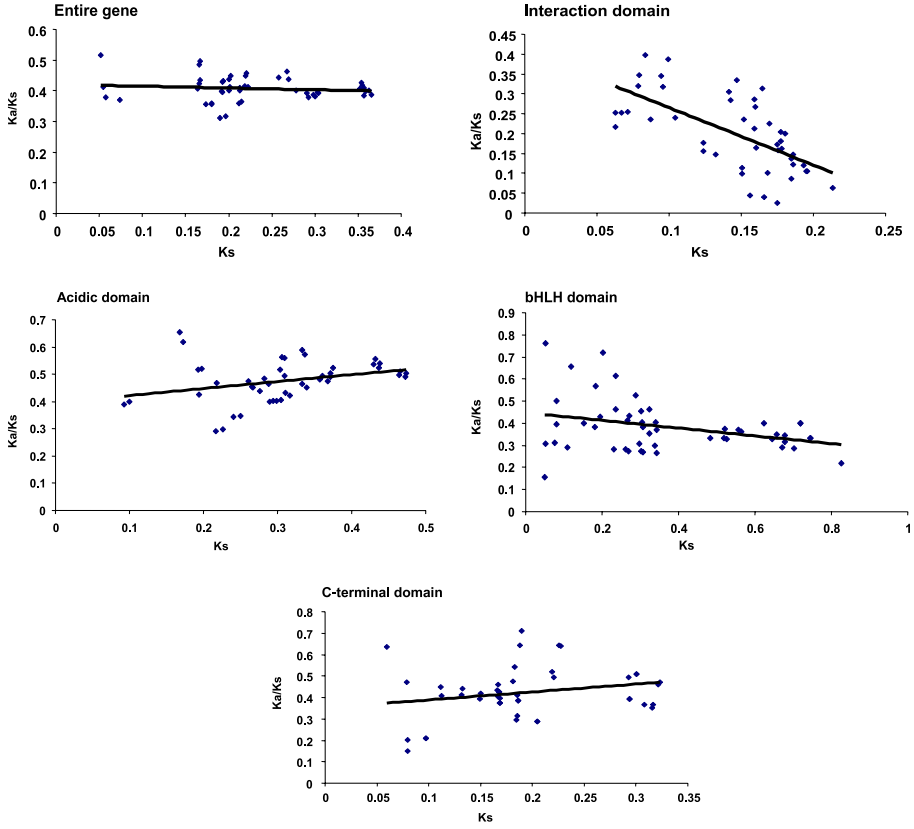

Fig. 6. Plots of K_a/K_s versus K_s for the entire gene, interaction domain, acidic domain, bHLH domain, and C-terminal domain for 11 *Cornus* species.

Table 4
Likelihood values and parameter estimates for the *myc*-like anthocyanin regulatory gene using codon-based substitution model of PAML

Model code	Ln	Tree length	K(ts/tv)	ω (K_a/K_s)	Estimates of parameters
M0 (one-ratio)	−5980.75	1.082	2.81	0.5132	$\omega = 0.5132$
M1 (neutral)	−5965.77	1.106	3.00	0.6251	$P_0 = 0.3749$ ($P_1 = 0.6251$)
M2 (selection)	−5962.58	1.110	2.88	0.5388	$P_0 = 0.00$, $P_1 = 0.4563$ ($P_2 = 0.5437$), $\omega_2 = 0.1518$
M3 (discrete)	−5962.56	1.111	2.88	0.5441	$P_0 = 0.00$, $P_1 = 0.5728$ ($P_2 = 0.4272$); $\omega_0 = 0.0102$, $\omega_1 = 0.1686$, $\omega_2 = 1.048$
M7 (beta)	−5962.97	1.107	2.87	0.5326	$p = 0.3042$ ($q = 0.2670$)
M8 (beta & ω)	−5962.56	1.111	2.88	0.5441	$p = 20.54$, $q = 99.00$; $P_0 = 0.5760$($P_1 = 0.4240$); $\omega = 1.050$

Ln, log likelihood value of NJ tree; ts, transition; tv, transversion; K_a, nonsynonymous substitution rate; K_s, nonsynonymous substitution rate. See Yang et al. (2000) for the definitions of parameters.

Table 5
Likelihood ratio test comparing models of variable ω ratios among sites

Comparisons	Log-likelihood values	Degree of freedom (v)	χ^2 distribution at v	LRT statistic
M1 vs M2	−5965.77 vs −5962.58	2	$P0.05 = 5.99$; $P0.01$=9.21	6.38*
M0 vs M3	−5980.75 vs −5962.56	4	$P0.01 = 13.28$; $P0.005 = 14.86$	36.38**
M7 vs M8	−5962.97 vs −5962.56	2	$P0.1 = 4.61$; $P0.05 = 5.99$	0.82

* Significant difference at $P < 0.05$ level.
** Significant difference at $P = 0.01$ level.

an adequate number of phylogenetically informative sites, and nucleotide sequences of this gene are informative of phylogeny in the dicot genus *Cornus* (Fig. 4). The sequences are more variable and contain a greater percentage of informative sites than do other nuclear (e.g., 26S rDNA, 18S rDNA) and chloroplast genes (e.g., *matK* and *rbcL*) applied in phylogenetic analyses of this genus. Our data also show that protein sequences of the gene could be potentially useful to resolve phylogenetic relationships at higher taxonomic levels (e.g.,

Table 6
The number and distribution of positive selection sites (with $\omega > 1$) detected using M3 and M8

Posterior probability	Total	Interaction domain	Acidic domain	bHLH domain	C-terminal domain
$P > 50\%$	241 (M3)	26 (M3)	126 (M3)	34 (M3)	36 (M3)
	240 (M8)	26 (M8)	125 (M8)	34 (M8)	36 (M8)
$P > 95\%$	27 (M3)	3 (M3)	17 (M3)	3 (M3)	3 (M3)
	27 (M8)	3 (M8)	17 (M8)	3 (M8)	3 (M8)
$P > 99\%$	6 (M3)	1 (M3)	4 (M3)	1 (M3)	0 (M3)
	5 (M8)	1 (M8)	4 (M8)	0 (M8)	0 (M8)

M3, discrete model; M8, beta & ω model.

Table 7
Pattern of positive selection sites with $P > 99\%$ ($\omega > 1$) suggested by codon-based substitution model

Position of positive selected sites	Amino acids variations	M3 model (discrete)	M8 model (beta & ω)	Domain position
89	A, V, S	Yes	Yes	Interaction
251	L, M, V, I	Yes	Yes	Acidic
312	S, F, I, L	Yes	Yes	Acidic
319	C, S, V	Yes	Yes	Acidic
378	M, I, T, V, A	Yes	Yes	Acidic
442	S, L, P	Yes	No	bHLH

family or above). Our phylogenetic analyses of protein sequences of *Cornus*, five additional dicot sequences, and four monocot sequences are highly congruent with the current view of angiosperm phylogeny (Figs. 2 and 3). Our data also show that alignable intron sequences among species within subgroups of *Cornus* can be useful for elucidating relationships among closely related species. Phylogenetic analyses using the entire genomic sequences including intron regions for 47 dwarf dogwood samples further demonstrate the phylogenetic utility of this gene at the intraspecific level (Fan et al., in preparation).

4.2. Rates and pattern of gene evolution

Regulatory genes have been shown to play an important evolutionary role in both plants and animals (reviews by Levine and Tjian, 2003; Papp et al., 2003; Purugganan, 1998, 2000). Major morphological changes have been linked to changes in regulatory genes rather than structural genes (King and Wilson, 1975; Wilson, 1975). Many studies have demonstrated that mutations in regulatory genes indeed cause dramatic shifts in morphology and functional activities (see reviews by Carroll, 1995; Doebley and Lukens, 1998; Kellogg, 2002; Palopoli and Patel, 1996; Purugganan, 1998; Simpson, 2002). Given that phenotypic changes may be the consequence of changes in gene expression, understanding rates and patterns of regulatory gene change within and among species is a critical step toward understanding biological evolution (Meiklejohn et al., 2003).

Because regulatory genes play important roles in the development and function of organisms, it is expected that regulatory genes might be highly constrained and evolve at relatively low rates. However, rapid mosaic evolution of regulatory genes was revealed in several recent studies in both animals (e.g., *Drosophila*) and plants (e.g., maize) with some regulatory domains (e.g., DNA-binding domains) evolving relatively slowly, and others changing quite rapidly. In general, sequence evolution for regulatory genes is often faster than that observed in structural genes (Alvarez-Buylla et al., 2000; Fridman and Zamir, 2003; Olsen et al., 2002; Purugganan, 1998; Ting et al., 1998).

Our study on the anthocyanin regulatory gene in *Cornus* indicated a mean ratio of nonsynonymous versus synonymous substitutions in *Cornus* greater than 0.4, a value significantly higher than that of most structural genes reported (e.g., 0.14 for a set of plant genes and 0.189 for 42 mammalian sequences; Li et al., 1985; Nei, 1987; Purugganan, 1998). Congruent with previous findings (Graur, 1985), the relationships between K_a/K_s and K_s in *Cornus* further indicated that the synonymous and nonsynonymous rates are significantly correlated, although in different fashions among domains (Fig. 6). This phenomenon suggests mosaic evolution of the gene and its domains.

Analyses of codon-based substitutions provide further evidence of mosaic evolution among the four functional domains of the *myc*-like anthocyanin regulatory gene-revealing considerable heterogeneity in rate of evolution of the gene among sites. The amino acid sequences of nine species of *Cornus* and five other dicots indicated one region that is well conserved: the nearly 200 residues of the interaction domain near the N-terminal region (13.9% variable sites within *Cornus*; 50% var-

iable sites within eudicots). However, in *Cornus*, the bHLH region has relatively more highly variable sites (58.6% variable sites within *Cornus*; 74.6% variable sites within eudicots) but with some conserved amino acid components. The region between these two domains is the acidic domain, which is less conserved (55.8% variable sites within *Cornus*; 72.3% within eudicots) and rich in negatively charged amino acids. The function of bHLH domain is believed to involve DNA-binding, as well as subunit dimerization activity of the *R* protein (Atchley and Fitch, 1997; Atchley et al., 2000; Murre et al., 1989). The function of the interaction domain involves transcriptional activation (Goff et al., 1992). Because of functional constraints, the interaction and bHLH domains are expected to evolve slowly and remain largely conserved in amino acid sequence. In maize, the conserved regions (most of these two domains) evolve at about 1.02×10^{-9} nonsynonymous substitutions/site/year, whereas the rest of the gene evolves approximately four times faster, at a significantly higher rate of 4.08×10^{-9} nonsynonymous substitutions/site/year (Purugganan and Wessler, 1994).

As expected, our data show that both K_a and K_s in the interaction and bHLH domains are lower than those in other domains, and the K_a/K_s is negatively related to K_s (Fig. 6). The significantly lower ratio of K_a/K_s found in the interaction domain compared to other domains suggests that this domain might be under the strongest functional constraint of all four functional domains of the gene. The interaction domain is divided into two sub-domains, interaction sub-domain I (aa 1–91 in *Cornus*) and interaction sub-domain II (aa 98–194 in *Cornus*) (Goff et al., 1992). Three tryptophan residues in sub-domain I (W-29, W-35, and W-47) and two tryptophans residues in sub-domain II (W-113 and W-141) are conserved across all taxa, including both dicots and monocots. These conserved tryptophan residues are thought to play an important role in forming a hydrophobic core for the *MYB*-like proteins (Anton and Frampton, 1988). Other conserved amino acids in these domains include the negatively charged aspartic and glutamic acids, which may form part of a hydrophilic surface (Ptashne, 1988). Transgenic studies using tobacco plants have shown that no pigmentation accumulation in flowers was observed with the deletion of sub-domain I and partial sub-domain II of *myc*-GP (Gong et al., 1999).

The bHLH domain is likely a key region of the *myc*-like/*R* regulatory protein. Within the bHLH domain, there are two highly conserved regions. The first region includes many basic residues that allow the helix–loop–helix to bind to DNA. The second is the HLH domain, including two amphipathic α-helices separated by a loop. This region is characterized by hydrophobic residues, which allow these proteins to interact and to form dimers (Murre et al., 1994). The bHLH domain includes 58 amino acids in *Cornus* with the key amino acids [e.g., glutamic acid (E) and arginine (R)], required for DNA binding, and hydrophobic residues [e.g., leucine (L)] at helix regions requiring the formation of a dimer, conserved among *Cornus* and other dicots (see Supplementary Materials).

Our phylogenetic analyses indicated a high percentage of variable sites among *Cornus* species and other dicots (e.g., 34 of 58 sites are variable within *Cornus*). However, the swap of amino acids mostly involves residues with similar chemical structures and/or charges, e.g., leucine (L)–valine (V), serine (S)–threonine (T), glutamic acid (E)–aspartic acid (D) (see Supplementary Materials), which would not severely affect the function of the protein. Furthermore, most of these substitutions are found in the loop region, which is the most variable region within the bHLH domain (Atchley et al., 1999). Previous studies show that chemically similar amino acids are known to be more interchangeable than chemically different ones due to redundancy of the genetic code and the effects of purifying selection (Gojobori et al., 1982).

Our data suggest that the acidic domain evolves most rapidly. This domain has the highest synonymous substitution rates and K_a/K_s (Fig. 5). Applying a codon-based model indicates that over half of sites potentially under positive selection are from the acidic domain (126 of 241 sites, 52.3%, Table 6). Furthermore, four of the six positive sites identified as positively selected with 99% probability also occur in the acidic domain. Amino acid substitutions at these four sites indicate that substitutions at three of them (sites 312, 319, and 378) involve changes between polar uncharged amino acids and non-polar amino acids (Table 7). Polar uncharged amino acids have partial positive or negative charges allowing their participation in chemical reactions, forming H-bonds, and association with water. Therefore, these sites might play a significant role in transactivation. Mutation analyses demonstrate that the acidic domain contains such transactivation sites (Gong et al., 1999). Previous comparison between myc-rp/gp in *Perilla* and *Delila* in *A. majus* found that *DELILA* exhibited higher transactivation activity than MYC-RP/GP (Gong et al., 1999). Amino acid alternation in this region may be the main reason for the different transactivation activities observed.

Color differences in petals are present among *Cornus* species. Among nine species examined, *C. suecica* has dark red or purple petals. *Cornus canadensis* has white petals and other species have yellowish/greenish petals. The observed higher rate of sequence evolution in the anthocyanin regulatory gene, particularly in the acidic domain, might contribute to the difference of flower colors. However, the association of sequence evolution and color variation is complex. We can detect no obvious correlation between sequence variation and flower color changes among different species. Therefore, further

developmental and genetic experiments are needed to test this hypothesis.

Acknowledgments

The authors thank the following people for providing different kinds of help: Brian Cassel for assistance with sequencing; Francesca Quattrocchio for providing the genomic sequences of *Petunia*-JAF13; Errol Strain for helping with data analyses using codon-based substitution models in PAML; Lisa David for extracting DNA for several samples. This study is partially supported by NSF grant DEB-0129069 to Q.-Y.X. and a Karling Graduate Student Research Award from the Botanical Society of America and the NSF Deep Gene program Travel Award to C.F.

Appendix. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ympev.2004.08.002.

References

Alvarez-Buylla, E.R., Liljegren, S.J., Pelaz, S., Gold, S.E., Burgeff, C., Ditta, G.S., Vergara-Silva, F., Yanofsky, M.F., 2000. MADS-box gene evolution beyond flowers: expression in pollen, endosperm, guard cells, roots and trichomes. Plant J. 24, 457–466.

Anton, N.J., Frampton, J., 1988. Tryptophans in myb proteins. Nature 336, 719.

Atchley, W.R., Fitch, W.M., 1997. A natural classification of the basic helix–loop–helix class of transcription factors. Proc. Natl. Acad. Sci. USA 94, 5172–5176.

Atchley, W.R., Terhalle, W., Dress, A., 1999. Positional dependence, cliques, and predictive motifs in the bHLH protein domain. J. Mol. Evol. 48, 501–516.

Atchley, W.R., Wollenberg, K.R., Fitch, W.M., Terhalle, W., Dress, A.W., 2000. Correlation among amino acid sites in bHLH protein domains: an information theoretic analysis. Mol. Biol. Evol. 17, 164–178.

Bailey, C.D., Doyle, J.J., 1999. Potential phylogenetic utility of the low-copy nuclear gene *Pistillata* in dicotyledonous plants: Comparison to nrDNA ITS and *trnL* intron in *Sphaerocardamum* and other Brassicaceae. Mol. Phylogenet. Evol. 13, 20–30.

Barrier, M., Baldwin, B.G., Robichaux, R.H., Purugganan, M.D., 1999. Interspecific hybrid ancestry of a plant adaptive radiation: allopolyploidy of Hawaiian silversword alliance (Asteraceae) inferred from floral homeotic gene duplications. Mol. Biol. Evol. 16, 1105–1113.

Barrier, M., Robichaux, R.H., Purugganan, M.D., 2001. Accelerated regulatory gene evolution in an adaptive radiation. Proc. Natl. Acad. Sci. USA 98, 10208–10213.

Bate, N.J., Rothstein, S.J., 1997. An *Arabidopsis myc*-like gene with homology to the anthocyanin regulatory gene *Delila* (Accession No. AF013465). Plant Physiol. 115, 315.

Carroll, S.B., 1995. Homeotic genes and the evolution of arthropods and chordates. Naturalist 376, 479–485.

Consonni, G., Viotti, A., Dellaporta, S.L., Tonelli, C., 1992. cDNA nucleotide sequence of *Sn*, a regulatory gene in maize. Nucleic Acids Res. 20, 373.

Consonni, G., Geuna, F., Gavazzi, G., Tonelli, C., 1993. Molecular homology among members of the *R* gene family from maize. Plant J. 3, 335–346.

Davis, R., Weintraub, H., Lasser, A., 1987. Expression of a single transfected cDNA converts fibroblasts into myoblasts. Cell 51, 1061–1067.

DePinho, R., Hatton, K., Tesfaye, A., Yancopoulos, G., Alt, F., 1987. The human *myc* gene family: structure and activity of *L-myc* and *L-myc* pseudogene. Genes and Dev. 1, 1311–1326.

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

Durbin, M.L., McCaig, B., Clegg, M.T., 2000. Molecular evolution of the chalcone synthase multigene family in the morning glory genome. Plant Mol. Biol. 42, 79–92.

Epperson, B., Clegg, M.T., 1987. Frequency-dependent variation for outcrossing rate among flower color morphs of *Ipomoea purpurea*. Evolution 41, 1302–1311.

Eyde, R.H., 1988. Comprehending *Cornus*: puzzles and progress in the systematics of dogwoods. Bot. Rev. 54, 233–351.

Fan, C., Xiang, Q.-Y., 2001. Phylogenetic relationships within *Cornus* (Cornaceae) based on 26S rDNA sequences. Am. J. Bot. 88, 1131–1138.

Fan, C., Xiang, Q.-Y., 2003. Phylogenetic analyses of Cornales based on 26S rDNA and combined 26S rDNA-*matK rbcL* sequence data. Am. J. Bot. 90, 1357–1372.

Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783–791.

Fridman, E., Zamir, D., 2003. Functional divergence of a synthetic invertase gene family in tomato, potato, and *Arabidopsis*. Plant Physiol. 131, 603–609.

Goff, S.A., Cone, K.C., Chandler, V.L., 1992. Functional analysis of the transcriptional activator encoded by the maize *B* gene: evidence for the direct functional interaction between two classes of regulatory proteins. Genes Dev. 6, 864–875.

Gojobori, T., Li, W.-H., Graur, D., 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. 18, 360–369.

Goldsbourough, A.P., Tong, Y., Yoder, J.I., 1996. *Lc* as a non-destructive visual reporter and transposition marker gene for tomato. Plant J. 9, 927–933.

Gong, Z., Yamagishi, E., Yamazaki, M., Saito, K., 1999. A constitutively expressed *myc*-like gene involved anthocyanin biosynthesis from *Perilla frutescens*: molecular characterization, heterologous expression in transgenic plants and transactivation in yeast cells. Plant Mol. Biol. 41, 33–44.

Goodrich, J., Carpenter, R., Coen, E.S., 1992. A common gene regulates pigmentation pattern in diverse plant species. Cell 68, 955–964.

Gottlieb, L.D., Ford, V.S., 1996. Phylogenetic relationships among the sections of *Clarkia* (Onagraceae) inferred from the nucleotide sequences of *PgiC*. Syst. Bot. 21, 45–62.

Graur, D., 1985. Amino acid composition and the evolutionary rates of protein-coding genes. J. Mol. Evol. 22, 53–62.

Holton, T.A., Cornish, E.C., 1995. Integrated control of seed maturation and germination programs by activator and repressor functions of viviparous-1 of maize. Genes. Dev. 9, 2459–2469.

Hu, J., Anderson, B., Wessler, S., 1996. Isolation and characterization of rice genes: evidence for distinct evolutionary paths in rice and maize. Genetics 142, 1021–1031.

Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules. In: Munro, H.N. (Ed.), Mammalian protein metabolism. Academic, New York, pp. 21–132.

Kellogg, E.A., 2002. Root hairs, trichomes and the evolution of duplicate genes. Trends Plant Sci. 6, 550–552.

King, J.L., Wilson, A.C., 1975. Evolution at two levels in humans and chimpanzees. Science 188, 107–116.

Kumar, S., Tamura, K., Jakobsen, I.B., Nei, M., 2001. MEGA2: Molecular Evolutionary Genetics Analysis software. Bioinformatics 17, 1244–1245.

Levine, M., Tjian, R., 2003. Transcription regulation and animal diversity. Nature 424, 147–151.

Li, W.-H., Wu, C.-I., Luo, C.-C., 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substation considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. 2, 150–174.

Liu, Y.-G., Whittier, R.F., 1995. Thermal asymmetric interlaced PCR: Automatable amplification and sequencing of insert end fragments from P1 and YAC clones for chromosome walking. Genomics 25, 674–681.

Liu, Y.-G., Mitsukawa, N., Oosumi, T., Whittier, R.F., 1995. Efficient isolation and mapping of *Arabidopsis thaliana* T-DNA insert junctions by thermal asymmetric interlaced PCR. Plant J. 8, 457–463.

Lloyd, A.M., Walbot, V., Davis, R.W., 1992. *Arabidopsis* and *Nicotiana* anthocyanin production activated by maize regulator *R* and *C1*. Science 258, 1773–1775.

Ludwig, S.R., Habera, L.F., Dellaporte, S.L., Wessler, S.R., 1989. *Lc*, a member of the maize *R* gene family responsible for tissue-specific anthocyanin production, encodes a protein similar to transcriptional activators and contains the *myc*-homology region. Proc. Natl. Acad. Sci. USA 86, 7092–7096.

Ludwig, S., Wessler, S.R., 1990. Maize *R* gene family: tissue-specific helix–loop–helix proteins. Cell 62, 849–852.

Marck, C., 1988. DNA Strider: a 'C' program for the fast analysis of DNA and protein sequences on the Apple Macintosh family of computers. Nucleic Acids Res. 16, 1829–1836.

Martin, C., Prescott, A., Machay, S., Bartlett, J., Vrijlandt, E., 1991. Control of anthocyanin biosynthesis in flowers of *Antirrhinum majus*. Plant J. 1, 37–49.

Mason-Gamer, R.J., Weil, C.F., Kellogg, E.A., 1998. Granule-Bound starch synthase: Structure, function, and phylogenetic utility. Mol. Biol. Evol. 15, 1658–1673.

Mathews, S., Sharrock, R.A., 1996. The phytochrome gene family in grasses (Poaceae): a phylogeny and evidence that grasses have a subset of the loci found in dicot angiosperms. Mol. Biol. Evol. 13, 1145–1150.

Mathews, S., Donoghue, M.J., 1999. The root of angiosperm phylogeny inferred from duplicate phytochrome genes. Science 286, 947–950.

Mathews, S., Tsai, R.C., Kellogg, E.A., 2000. Phylogenetic structure in the grass family (Poaceae): evidence from the nuclear gene phytochrome B. Am. J. Bot. 87, 96–107.

Meiklejohn, C.D., Parsch, J., Ranz, J.M., Hartl, D.L., 2003. Rapid evolution of male-biased gene expression in *Drosophila*. Proc. Natl. Acad. Sci. USA 100, 9894–9899.

Mol, J., Grotewold, E., Koes, R., 1998. How genes paint flower and seeds. Trends Plant Sci. 3, 212–217.

Morgan, D.R., Soltis, D., 1993. Phylogenetic relationships among members of Saxifragaceae sensu lato based on *rbcL* sequence data. Ann. MO. Bot. Gard. 80, 631–660.

Murre, C., McCaw, P.S., Baltimore, D., 1989. A new DNA binding and dimerization motif in immunoglobulin enhancer binding, *daughterless*, *MyoD*, and *myc* proteins. Cell 56, 777–783.

Murre, C., Bain, G., van Dijk, M.A., Engel, I., Furnari, B.A., Massari, M.E., Mathews, J.R., Quong, M.W., Rivera, R.R., Stuiver, M.H., 1994. Structure and function of helix–loop–helix proteins. Biochim. Biophys. Acta 1218, 129–135.

Nei, M., 1987. Molecular Evolutionary Genetics. Columbia University Press, New York.

Nei, M., Gojobori, T., 1986. Simple method for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3, 418–426.

Nielsen, R., Yang, Z., 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148, 929–936.

Nishimoto, Y., Ohnishi, O., Hasegawa, M., 2003. Topological incongruence between nuclear and chloroplast DNA trees suggesting hybridization in the urophyllum group of the genus *Fagopyrum* (Polygonaceae). Genes. Genet. Syst. 78, 139–153.

Olsen, K.M., Schaal, B.A., 1999. Evidence on the origin of cassava: Phylogeography of *Manihot esculenta*. Proc. Natl. Acad. Sci. USA 96, 5586–5591.

Olsen, K.M., Womack, A., Garrett, A.R., Suddith, J.I., Purugganan, M.D., 2002. Contrasting evolutionary forces in the *Arabidopsis thaliana* floral developmental pathway. Genetics 160, 1641–1650.

Palopoli, M.F., Patel, N., 1996. Neo-Darwinian developmental evolution – can we bridge the gap between pattern and process. Curr. Opin. Genet. Dev. 6, 502–508.

Papp, B., Pál, C., Hurst, L.D., 2003. Evolution of *cis*-regulatory elements in duplicated genes of yeast. Trends Genet. 19, 417–422.

Perrot, G.H., Cone, K.C., 1989. Nucleotide sequence of the maize R-S gene. Nucleic Acids Res. 17, 8003.

Posada, D., Crandall, K.A., 1998. Modeltest: testing the model of DNA substitution. Bioinformatics 14, 817–818.

Ptashne, M., 1988. How eukaryotic transcriptional activators work. Nature 335, 683–689.

Purugganan, M.D., 1998. The molecular evolution of development. BioEssays 20, 700–711.

Purugganan, M.D., 2000. The molecular population genetics of regulatory genes. Mol. Ecol. 9, 1451–1461.

Purugganan, M.D., Wessler, S.R., 1994. Molecular evolution of the plant *R* regulatory gene family. Genetics 138, 849–854.

Quattrocchio, F., Wing, J.F., Leppen, H.T.C., Mol, J.N.M., Koes, R.E., 1993. Regulatory genes controlling anthocyanin pigmentation are functionally conserved among plant species and have distinct sets of target genes. Plant Cell 5, 1497–1512.

Quattrocchio, F., Wing, J.F., van der Woude, K., Mol, J.N.M., Koes, R., 1998. Analysis of bHLH and MYB domain proteins: species-specific regulatory differences are caused by divergent evolution of target anthocyanin genes. Plant J. 13, 475–488.

Radicella, J.P., Turks, D., Chandler, V.L., 1991. Cloning and nucleotide sequence of a cDNA encoding *B-peru*, a regulatory protein of the anthocyanin pathway from maize. Plant Mol. Biol. 17, 127–130.

Sang, T., 2002. Utility of low-copy nuclear gene sequences in plant phylogenetics. Crit. Rev. Biochem. Mol. Biol. 27, 121–147.

Sang, T., Donoghue, M.J., Zhang, D., 1997. Evolution of alcohol dehydrogenase genes in Peonies (*Paeonia*): phylogenetic relationships of putative nonhybrid species. Mol. Biol. Evol. 14, 994–1007.

Sang, T., Zhang, D., 1999. Reconstructing hybrid speciation using sequences of low-copy nuclear genes: hybrid origins of five *Paeonia* species based on *Adh* gene phylogenies. Syst. Bot. 24, 148–163.

Simpson, P., 2002. Evolution of development in closely related species of flies and worms. Nat. Rev. Genet. 3, 907–917.

Small, R.L., Ryburn, J.A., Cronn, R.C., Seelanan, T., Wendel, J.F., 1998. The tortoise and the hare: choosing between noncoding plastome and nuclear *Adh* sequences for phylogeny reconstruction in recently diverged plant group. Am. J. Bot. 85, 1301–1315.

Small, R.L., Wendel, J.F., 2000. Phylogeny, duplication, and intraspecific variation of *Adh* sequences in new world diploid cottons (*Gossypium* L., Malvaceae). Mol. Phylogenet. Evol. 16, 73–84.

Soltis, D.E., Soltis, P.S., 1997. Phylogenetic relationships among Saxifragaceae sensu lato: a comparison of topologies based in 18S rDNA and *rbc* L sequences. Am. J. Bot. 84, 504–522.

Soltis, D.E., et al., 2000. Angiosperm phylogeny inferred from 18S rDNA, *rbcL*, and *atpB* sequences. Biol. J. Linn. Soc. 133, 381–461.

Spelt, C., Quattrocchio, F., Mol, J.N.M., Koes, R., 2000. Anthocyanin1 of *Petunia* encodes a basic helix–loop–helix protein that directly activates transcription of structural anthocyanin genes. Plant Cell 12, 1619–1631.

Stapleton, A., 1992. Ultraviolet radiation and plants: burning questions. Plant Cell 4, 1353–1358.

Swofford, D.L., 2002. PAUP: phylogenetic analysis using parsimony, version 4.0b10. Sinauer Associates, Sunderland, MA.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The Clustal X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 24, 4876–4882.

Ting, C.T., Tsaur, S.-C., Wu, M.-L., Wu, C.-I., 1998. A rapidly evolving homeobox at the site of a hybrid sterility gene. Science 282, 1501–1504.

Wilson, A.C., 1975. Evolutionary importance of gene regulationStadler Symposium, 7. University of Missouri, Columbia, MO, pp. 117–134.

Xiang, Q.-Y., Soltis, D.E., Morgan, D.R., Soltis, P.S., 1993. Phylogenetic relationships of *Cornus* L. sensu lato and putative relatives inferred from *rbcL* sequence data. Ann. MO. Bot. Gard. 80, 723–734.

Xiang, Q.-Y., Soltis, D.E., Soltis, P.S., 1998. Phylogenetic relationships of Cornaceae and close relatives inferred from *matK* and *rbcL* sequences. Am. J. Bot. 85, 285–297.

Yang, Z., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS 13, 555–556.

Yang, Z., Nielsen, R., Goldman, N., Pedersen, A.-M., 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155, 431–449.