ABSTRACT

Fan, Chuanzhu. Molecular Phylogeny and Evolution of Dogwoods. (Under the direction of Dr. Jenny Qiu-Yun Xiang)

Dogwoods consist of morphologically diverse plants, and taxonomic circumscription and phylogenetic relationships of dogwoods have long been controversial. My dissertation study has two major goals: 1) elucidate phylogenetic relationships in the dogwoods group using nuclear DNA sequences, and 2) investigate the sequence evolution and its morphological link of the *myc*-like anthocyanin regulatory gene and explore the phylogenetic utility of the gene in dogwoods. Phylogenetic relationships within *Cornus* and Cornales (*Cornus* and related genera and families) were previously investigated using chloroplast DNA sequence data in several studies, but these still remained incompletely resolved. I used nuclear 26S rDNA sequences to further elucidate relationships within the group and to corroborate previously published phylogenetic hypotheses based on cpDNA and morphological data. Phylogenetic analyses of 26S rDNA sequence data (~3.4 kb) in combination with sequences of chloroplast genes *rbcL* and *matK*, suggest that the aquatic enigmatic genus, *Hydrostachys* from southern Africa, is sister to the remainder of Cornales among which *Cornus* and *Alangium* are sisters, nyssoids (*Nyssa*, *Camptotheca*, and *Davidia*) and mastixioids (*Mastixia* and *Diplopanax*) are sisters, and Hydrangeaceae and Loasaceae are sisters. These relationships, except the placement of *Hydrostachys*, are consistent with previous findings from analyses of *matK-rbcL* sequence data. Within *Cornus*, the dwarf dogwoods (subg. *Arctocrania*) are the sister of the big-bracted dogwoods (subg. *Cynoxylon* and subg. *Syncarpea*). This clade is, in turn, sister to the cornelian cherries (subg. *Cornus* and

subg. *Afrocrania*). This large red-fruited clade is sister to a clade consisting of the blue- or white-fruited species (subg. *Mesomora*, subg. *Kraniopsis,* and subg. *Yinquania*). Within latter clade, *C. oblonga* (subg. *Yinquania*) is sister to the remainder, and subg. *Mesomora* is sister to subg. *Kraniopsis*. These relationships are congruent with those suggested by cpDNA sequences data, but with greater statistical support when the cpDNA data are combined with nuclear DNA sequences.

To accomplish the second goal, the entire sequences (~4 kb) of the *myc*-like anthocyanin regulatory gene were sequenced for nine species of *Cornus* representing all four major clades of the genus and 47 samples of the dwarf dogwood species complex. Our phylogenetic analyses of sequences indicate that the *myc*-like anthocyanin regulatory gene is phylogenetically useful at different taxonomic levels depending on the data set (nucleotide vs. protein sequences) and regions (exons vs. introns) applied. Amino acid sequences are useful to resolve relationships among families of flowering plants, whereas nucleotide sequences from the coding region are useful to resolve relationships among subgroups of *Cornus* and DNA sequences of the entire gene are informative among closely related species within subgroups of *Cornus*. Sequence evolution of the gene was examined using a codon-based substitution model and population genetic methods. All results indicate accelerated sequence evolution of this gene at both interspecific and intraspecific levels. Mosaic evolution and heterogeneous rates in DNA sequence were detected among the four functional domains and among sites. The interaction domain, involving an important function, has the lowest ratio of nonsynonymous and synonymous substitution rate, suggesting the strongest evolutionary constraint. The acidic domain evolves most rapidly among four domains. In the bHLH domain, the key residues are

conserved among all species examined, although its evolutionary rate is faster than that of the interaction domain. Most changes in this domain occur in loop region and among amino acids with similar chemical features. Furthermore, sites under positive selection were detected. Nucleotide diversity and neutrality tests of DNA sequences suggested that an excess of low-frequency polymorphisms and an excess of replacement substitutions exist within the dwarf dogwoods complex. Positive selection and/or recent, rapid population expansion are the main forces acting on the accelerated gene evolution and result in an excess of low-frequency polymorphism. Significant correlation between petal color and amino acid sequence variation in the dwarf dogwoods complex was detected by ANOVA analyses using GLM model. Substitutions at three amino acid sites are significantly associated with petal color, which suggests that the mutation at these sites may result in changes of protein function, and are thus responsible for the color change in flowers. Extensive gene flow, gene recombination, and introgression were also detected within the dwarf dogwoods complex, suggesting a dynamic evolutionary process within the dwarf dogwoods.

MOLECULAR PHYLOGENY AND EVOLUTION OF DOGWOODS

by

CHUANZHU FAN

A dissertation submitted to the Graduate Faculty of North Carolina State University

in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

DEPARTMENT OF BOTANY

Raleigh, NC

November, 2003

APPROVED BY:

Dr. (Jenny) Qiu-Yun Xiang
Chair of Advisory Committee

Dr. Thomas R. Wentworth
Co-chair of Advisory Committee

Dr. Michael D. Purugganan

Dr. Brian M. Wiegmann

BIOGRAPHY

Chuanzhu was born and raised at a fishing village near the Yellow Sea, in the Northeast of China. He graduated from Northeast Normal University, Changchun, China, to earn a Bachelor degree of Science in Biology in July, 1989. He immediately entered graduate school at the Chinese Academy of Agricultural Sciences, Beijing, to pursue a master's degree. He was awarded the Master of Science in Plant Genetics, in June, 1992. He worked as an assistant research professor at the research institute where he earned his master's degree; after two years, he was promoted to director of the department of seed storage in the Institute of Crop Germplasm Resources. After seven years' working experience as research assistant professor, he decided to pursue his Ph.D. in the USA. He began his doctoral studies at Idaho State University in August, 1999. A year later, he transferred to North Carolina State University to continue his Ph.D. in North Carolina. Chuanzhu has accepted a post doctoral research position at the University of Chicago. He will work with Dr. Manyuan Long to study gene origination and evolution in fruit flies and other organisms. He married to his wife, Xingji (Cindy), in 1992. Their first son, Jiachen (Jason) was born in 1999, and second son, Frederick, was landed in 2002.

ACKNOWLEDGEMENTS

Many people have made my stay in the Botany Department, NCSU, a productive and enjoyable time. My deepest gratitude goes to my advisor, Dr. Jenny Xiang, for her insight, support, guidance, patience, and enthusiasm. Many thanks go to my committee member, Dr. Thomas R. Wentworth, Dr. Michael D. Purugganan, Dr. Brian M. Wiegmann, for their guidance, helpful discussions, suggestions, and encouragements. My appreciation also goes to the members of Xiang's Laboratory, David Thomas, Kathy McKeown, Wenheng Zhang, and Jennifer Modliszewski, for their help and expertise. I also want to thank the following people who gave me help with experiments and data analyses: O'Leary-Jepsen at Idaho State University and Art Johnson at the DNA sequencing and mapping facility of North Carolina State University for running some sequencing gels; Brian Cassel for assistance sequencing; Francesca Quattrocchio for providing the genomic sequence of *Petunia-jaf13*; Jim Qi for sample collection in the Alaska field trip; Nina Gardner for dwarf dogwoods DNA extraction and morphological identification; Errol Strain for helping data analyses using codon-based substitution models in PAML; Xi Chen for helping ANOVA test.

Finally, my special thanks and dedication go to my wife, Xingji (Cindy) Zhu, to my sons, Jason and Fred, who always give me inspiration and encouragement.

This work was supported by Faculty Research Grant from Idaho State University and North Carolina State University, NSF grant DEB-0129069 to Q-Y.X.; Karling Graduate Student Research Award from Botanical Society of America. I also want to thank North Carolina Plant Molecular Biology Consortium and Deep Gene Network for travel support during my studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

I. Phylogenetic Analyses of Cornales Based on 26S rDNA and Combined 26S rDNA-*matK-rbcL* Sequence Data

II. Phylogenetic Relationships within *Cornus* (Cornaceae) Based on 26S rDNA Sequences

III. Heterogeneous Evolution of the *myc*-like Anthocyanin Regulatory Gene in *Cornus* (Cornaceae)

LIST OF ABBREVIATIONS

aa - amino acid
ANOVA - analysis of variance
AP3 - *APETALA3*
bHLH - basic helix loop helix
bp - base pair
CAL - *CAULIFLOWER*
CC - conserved core
CI - consistency index
cpDNA - chloroplast DNA
DI water - deionized water
DMSO - dimethyl sulfoxide
DNA - deoxyribonucleic acid
dNTP - dinucleotide tri-phosphate
DS - double-stranded
ES - expansion segments
GLM - general linear model
GTR - general time reversible
ILD - incongruence length difference
ITS - inter transcribed spacer
Ka - nonsynonymous substitution
kb - kilobase
Ks - synonymous substitution
LRT - likelihood-ratio test
MEGA - molecular evolutionary genetic analysis
matK- maturase K
ML - maximum likelihood
MP - maximum parsimony
MYA - million years ago
PAML - phylogenetic analysis by maximum likelihood
PAUP - phylogenetic analysis using parsimony
PCR - polymerase chain reaction
PEG - polyethylene glycol
PI - *PISTILLTA*
rbcL - ribulose bisphosphate carboxylase large subunit
rDNA - ribosomal DNA
RI - retention index
SD - standard deviation
TAIL PCR - thermal asymmetric interlaced PCR
TBR - tree-bisection-reconnection

Chapter I

Phylogenetic analyses of Cornales based on 26S rDNA and combined 26S rDNA-*matK*-*rbcL* sequence data[1]

Chuanzhu Fan[2] and (Jenny) Qiu-Yun Xiang

Department of Botany, North Carolina State University, Campus Box 7612, Raleigh, North Carolina 27695-7612, USA

Published in *American Journal of Botany*, 2003, 90(9): 1357-1372

Chuanzhu did all of the work reported in this paper, but Dr. Jenny Xiang provided scientific advice and guidance.

[1]Manuscript received _____; revision received _____.[2]Author for reprint requests (email: cfan3@unity.ncsu.edu).

The authors thank O'Leary-Jepsen and the Molecular Core Facility at Idaho State University and the DNA sequencing and mapping facility at North Carolina State University for running some sequencing gels; Brian Wiegmann at NCSU for sharing the sequencing facility and for helpful comments on the manuscript; the following people for providing DNA samples or vouchered leaf samples for DNA extraction: D. E. Boufford - *Cornus officinalis*, C. Brochmann - *Cornus suecica*, M. W. Chase - *Grubbia*, Y-F Deng – *Alangium*, R. Fernandez Nava - *Cornus disciflora*, L. Hufford - *Mentzelia* and *Petalonyx*, J. Li – *Diplopanax*, Z. Murrel - *Cornus volkensii*, and N. Ronsted, R. Olmstead, B. Bremer – *Hydrostachys*; R. Olmstead and one anonymous reviewer for their insightful comments. This study is supported by NSF grant DEB-0129069.

Abstract: Nuclear 26S rDNA sequences were used to corroborate and test previously published *matK-rbcL*-based hypotheses of phylogenetic relationships in Cornales. Sequences were generated for 53 taxa including *Alangium*, *Camptotheca*, *Cornus*, *Curtisia*, *Davidia*, *Diplopanax*, *Mastixia*, *Nyssa*, and four families: Grubbiaceae, Hydrangeaceae, Hydrostachyaceae, and Loasaceae. Fifteen taxa from asterids were used as outgroups. The 26S rDNA sequences were initially analyzed separately and then combined with *matK-rbcL* sequences, using both parsimony and maximum likelihood methods. Eight strongly supported major clades were identified within Cornales by all analyses: *Cornus, Alangium*, nyssoids (*Nyssa*, *Davidia*, and *Camptotheca*), mastixioids (*Mastixia* and *Diplopanax*), Hydrangeaceae, Loasaceae, *Grubbia-Curtisia*, and *Hydrostachys*. However, relationships among the major lineages are not strongly supported in either 26S rDNA or combined 26S rDNA-*matK-rbcL* topologies, except for the sister relationships between *Cornus* and *Alangium*, and between nyssoids and mastixioids in the tree from combined data. Discrepancies in relationships among major lineages, especially the placement of the long branched *Hydrostachys*, were found between parsimony and maximum likelihood trees in all analyses. Incongruence between the 26S rDNA and *matK-rbcL* data sets was suggested, where Hydrangeaceae was found to be largely responsible for the incongruence. The long branch of *Hydrostachys* revealed in previous analyses was reduced significantly with more sampling. Maximum likelihood analysis of combined 26S rDNA-*matK-rbcL* sequences suggested that *Hydrostachys* might be sister to the remainder of Cornales, *Cornus-Alangium* are sisters, nyssoids-mastixioids are sisters, and Hydrangeaceae-Loasaceae are sisters, consistent with previous analyses of *matK-rbcL* sequence data.

Key words: 26S rDNA; Cornales; Grubbiaceae; Hydrostachyaceae; incongruence; long-branch attraction; *matK-rbcL*; phylogenetics.

INTRODUCTION

The order Cornales consists of a diversity of plants from herbs to shrubs and trees. Taxonomic circumscription of Cornales and phylogenetic relationships within Cornales have long been controversial. Early in 1898, Harms recognized a broadly defined Cornaceae (*Alangium*, *Aucuba*, *Camptotheca*, *Cornus*, *Corokia*, *Curtisia*, *Davidia*, *Garrya*, *Griselinia*, *Helwingia*, *Kaliphora*, *Melanophylla*, *Nyssa*, and *Toricellia*) divided among seven subfamilies in the order Umbelliflorae. Many of the genera originally placed in Cornaceae by Harms (1898) were later treated as distinct families placed in Cornales or moved to other orders by various authors (Table 1). Several recent phylogenetic analyses using chloroplast DNA data suggested a Cornales clade that differs from all previous traditional concepts (Chase et al., 1993; Olmstead et al., 1993; Xiang et al., 1993, 1998; Xiang and Soltis, 1998; Xiang, 1999; Albach et al., 2001a, b). This new Cornales clade contains several genera that have been previously placed in Cornaceae (*Alangium*, *Camptotheca*, *Cornus*, *Curtisia*, *Davidia*, *Diplopanax*, *Mastixia*, *Nyssa*), and four other families (Grubbiaceae, Hydrangeaceae, Hydrostachyaceae, and Loasaceae). Nine genera (*Aralidium*, *Aucuba*, *Corokia*, *Garrya, Griselinia*, *Helwingia*, *Kaliphora*, *Melanophylla*, and *Toricellia*) previously placed in Cornaceae (or Cornales) by some authors, were found to be related to non-cornalean asterids in those studies. This molecular-based circumscription of Cornales was followed by the Angiosperm Phylogeny Group (APG) in 1998.

The placement of Hydrangeaceae and Loasaceae within the Cornales clade was found in all large scale phylogenetic analyses using single or multiple molecular data sets (e.g., Downie and Palmer, 1992; Morgan and Soltis, 1993; Olmstead et al., 1993, 2000; Xiang,

1999; Soltis et al., 2000; Albach et al., 2001b). The two enigmatic monogeneric African families, Hydrostachyaceae and Grubbiaceae, first appeared in the cornalean clade in analyses of *rbcL* sequences for Loasaceae and Ebenales, respectively (Morton et al., 1996; Hempel et al., 1995). The affinity of these two families to Cornales was supported in further analyses for Cornales, asterids, eudicots, and angiosperms using chloroplast and nuclear genes (e.g., Xiang, 1999; Savolainen et al., 2000a, b; Soltis et al., 2000; Albach et al., 2001b; Xiang et al., 2002). Relationships within Cornales were not completely resolved in these previous studies, especially the placement of Hydrostachyaceae, for which there are few morphological features linking it to other cornalean taxa. Further, this family was always linked by an extremely long branch in the phylogenetic trees and its placement within Cornales was never strongly supported by bootstrap analyses (Hempel et al., 1995; Xiang, 1999; Soltis et al., 2000; Albach et al., 2001a; Xiang et al., 2002), and sometimes it was even placed outside of Cornales (Hufford et al., 2001). In the most recent analysis of *rbcL* and *matK* sequences for Cornales (Xiang et al., 2002), four major clades were identified: *Cornus*-*Alangium*, nyssoids-mastixioids (*Camptotheca*, *Davidia*, *Nyssa*, *Diplopanax*, and *Mastixia*), Hydrangeaceae (including *Hydrostachys*)-Loasaceae, and *Curtisia*-*Grubbia*, with the first two clades being sisters, which in turn are sister to the third clade. However, these relationships and the placement of *Hydrostachys* were not strongly supported by bootstrap analyses.

In the present study, we collected a new molecular data set, nuclear 26S rDNA sequences, to further elucidate phylogenetic relationships within Cornales. 26S rDNA sequences have been used to reconstruct phylogenetic relationships at various taxonomic levels of seed plants (e.g., Mishler et al., 1994; Ro et al., 1997; Kuzoff et al., 1998; Soltis and

Soltis, 1998; Stefanovic et al., 1998; Ro et al., 1999; Ashworth, 2000; Chanderbali et al., 2001; Fan and Xiang, 2001; Fishbein et al., 2001; Neyland, 2001; Simmons et al., 2001; Soltis et al., 2001; Nickrent et al., 2002; and Zanis et al., 2002). As 26S rDNA contains rapidly evolving expansion segments (ES) and conserved core (CC) regions (Clark et al., 1984; Dover and Flavell, 1984; Flavell, 1986), another goal of this study is to characterize the two regions (ES and CC) and evaluate their phylogenetic utilities in Cornales.

Long-branch attraction has long been recognized as a potential problem of parsimony analysis (Felsenstein, 1978; Swofford et al., 1996), whereas ML (maximum likelihood) methods incorporating appropriate substitution models may overcome this problem (Swofford et al., 1996). Given that *Hydrostachys* has been identified having extremely long branches in all previous studies, we analyzed our data using parsimony and ML methods to see how the two methods perform differently regarding its placement.

MATERIALS AND METHODS

Sampling-Fifty-three taxa from Cornales were sampled for the 26S rDNA sequencing study, including *Alangium*, *Camptotheca*, *Cornus*, *Curtisia*, *Davidia*, *Diplopanax*, *Mastixia*, and *Nyssa*, two species of *Grubbia*, 12 genera from Hydrangeaceae, three genera of Loasaceae, and seven taxa of *Hydrostachys* representing six species. The sampling for *Alangium*, *Hydrostachys,* and *Mastixia* is broader than all the previous studies (e.g. Xiang et al., 1998, 2002), including more species from these genera in an attempt to break potential long branches. A broad range of taxa (15 species) from Ericales and euasterids were chosen as outgroups. Among these, five were from Ericales (*Fouquieria columnaris*:

Fouquieriaceae, *Halesia diptera*: Halesiaceae, *Sarracenia*: Sarraceniaceae, *Shortia galacifolia*: Diapensiaceae, and *Styrax japonicus*: Styracaceae); four were sampled from Euasterids I including one species from Garryales *(Garrya elliptica*: Garryaceae), two species from Lamiales (*Myoporum mauritianum*: Scrophulariaceae and *Veronica anagallis-aquatica*: Veronicaceae), one species from Solanales (*Solanum lycopersicum*: Solanaceae); and six were from Euasterids II including three species from Apiales (*Apium graveolens*: Apiaceae, *Panax quinquefolius*: Araliaceae, and *Petroselinum crispum*: Apiaceae), two species from Asterales (*Corokia cotoneaster*: Argophyllaceae and *Tragopogon dubius*: Asteraceae), and one species from Aquifoliales (*Ilex opaca*: Aquifoliaceae). The 26S rDNA sequences for all outgroups were downloaded from GenBank except those for *Sarracenia* and *Shortia*, whose sequences were generated in this study. Complete list of taxa, voucher and GenBank accession numbers has been archived at the Botanical Society of America website (http://ajbsupp.botany.org/v90; Table 2).

DNA extraction-Most genomic DNAs used in this study were isolated for previous *rbcL* and *matK* sequencing studies. The new DNAs were extracted for *Alangium chinense*, *Alangium kurzii*, *Cornus disciflora*, *Hydrostachys polymorpha*, *Hydrostachys* spp., *Mentzelia decapetala*, *Mastixia eugenioides*, *Mastixia pentandra subsp. chinensis, Petalonyx parryi*, and *Shortia galacifolia* from dried leaves using the modified CTAB method of Cullings (1992) with modifications described in Xiang et al. (1998).

Gene amplification-The entire 26S rDNA (approximately 3.3 kb) was successfully amplified from total DNA aliquots via a single PCR (polymerase chain reaction) run for a few taxa using the forward primer N-nc26S1 (5'-CGACCCCAGGTCAGGCG-3') and the reverse primer 3331rev (5'-ATCTCAGTGGATCGTGGCAG-3') following Kuzoff et al.

(1998) with slight modifications. For most species, the entire 26S rDNA sequence was amplified in two segments using primers N-nc26S1 with 1449rev (5' - ACCCATGTGCAAGTGCCGTT - 3') and N-nc26S5 (5' - CGTGCAAATCGTTCGTCT - 3') or N-nc26S6 (5' - TGGTAAGCAGAACTGGCG - 3') with 3331rev. Our PCR reactions are described in Fan and Xiang (2001).

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). The purified DS DNA products were used as the templates for sequencing using the ABI PRISM dRhodamine Terminator Cycle Sequencing Ready Reaction Kit (Applied Biosystems, Foster City, California, USA). Cycle-sequencing reactions (10 μL) were prepared by combining 2 μL terminator ready reaction mix, 2 μL sequencing buffer (200 mmol/L Tris-ph8.0, 5mmol/L $MgCl_2$), 0.6 μL primer (5 μmol/L), 0.5 μL of 200 ng/μL cleaned PCR product, 0.5 μL dimethyl sulfoxide (DMSO), and 4.4 μL (deionized) DI water. Addition of 0.5 μL DMSO to the sequencing reactions resulted in cleaner sequences. Sixteen sequencing primers (N-nc26S1, N-nc26S3, N-nc26S4, N-nc26S5, N-nc26S6, N-nc26S8, N-nc26S10, N-nc26S12, N-nc26S14, 268rev, 641rev, 950rev, 1449rev, 2134rev, 2782rev, and 3331rev) described in Kuzoff et al. (1998), were used in different combinations to obtain the complete sequence of 26S rDNA. Cycle-sequencing was conducted on a PTC-100 Programmable Thermal Controller (MJ Research, Inc. Watertown, MA, USA) as follows: 25 cycles of 96ºC for 30 sec, 50ºC for 15 sec, and 60ºC for 4 min.

Products of cycle-sequencing were cleaned using ethanol/sodium acetate precipitation (ABI applied Biosystems, Foster City, California 94404, USA) with an additional 95% ethanol wash. The cleaned sequencing products were analyzed on an ABI-377 automated

sequencer (Applied Biosystems, Foster City, California 94404, USA). The sequence chromatogram output files for all samples were checked and edited base by base manually before being aligned. For a few species (*Cornus controversa*, *Cornus sessilis*, *Curtisia*, and *Hydrostachys*), the above sequence primers did not yield complete sequences due to sequence divergence in some primer regions. Four new primers, 1227F (5' - GAACCCACAAAGGGTGTTGGTCG - 3') and 1793R (5' - CGCGACGTGCGGTGCTCTTCCAG - 3') for *C. controversa* and *C. sessilis* 1951F (5' - TTCGGGAAAAGGATTGGCTCTGAGG - 3') and 2857R (5' – GTGGTAACTTTTCTGACACCTCTAG - 3') for *Curtisia* and *Hydrostachys* were designed to solve this problem.

Parsimony analysis - The 26S rDNA sequences were initially aligned using ClustalX (Thompson et al., 1997), and then adjusted manually. The aligned sequences consist of 68 taxa and 3430 bp (base pair) with small gaps (1-10 bp). The ES and CC regions of 26S rDNA were identified and located according to the coordinates for the expansion segments in the sequences of *Oryza sativa* (see Kuzoff et al., 1998) and *Cornus* (Fan and Xiang, 2001). The data matrix was analyzed with both parsimony and ML methods using PAUP* 4.0b10 (Swofford, 2002). For parsimony analysis, gaps were coded as missing data. Heuristic searches were performed using the MULPARS option with characters equally weighted, character states unordered, random taxon addition with 1000 replicates, tree-bisection-reconnection (TBR) branch-swapping. To evaluate clade support, 10000 replicates of bootstrap analysis (Felsenstein, 1985) were performed using fast heuristic search and TBR branch-swapping. In addition to analyses of the entire 26S rDNA sequences, ES and CC

regions were also analyzed separately using parsimony to compare the relative phylogenetic utilities of the two regions.

Model test and maximum likelihood analysis-In order to find the appropriate substitution models for ML analyses of the 26S rDNA sequence data, *matK-rbcL* sequence data, and the combined 26S rDNA-*matK-rbcL* data, model searching was performed using the software ModelTest (Posada and Crandall, 1998). ML analyses were subsequently conducted using the best model identified and parameter values estimated from ModelTest. For all ML analyses, heuristic searches were conducted using random taxon addition with ten replicates. Due to the extremely enormous amount of time required for bootstrap analyses of these large data sets (26S rDNA:3430 bp, and 26S rDNA-*matK-rbcL*: 6348 bp) using ML methods, we used neighbor joining bootstrap analysis employing ML distance to approximate the bootstrap supports for the ML trees. The same substitution model and parameters used in the ML analysis were used in the ML distance estimation. Ten thousand bootstrap replicates were conducted.

Incongruence test and combined data analysis. A combined data matrix of 26S rDNA-*matK-rbcL* including 42 taxa with sequences available for at least two of the three genes was constructed for a total evidence analysis. This matrix contains one species from Grubbiaceae, two from Hydrostachyaceae, four from Loasaceae, 13 from Hydrangeaceae, all 14 traditional cornalean genera, and eight outgroups. The aligned sequences contain a total of 6348 bp for each taxon, among which 3407 bp were from 26S rDNA, 1504 bp from *rbcL*, and 1437 bp from *matK*. Incongruence length difference test (ILD; Mickevich and Farris, 1981; Farris et al., 1994) was performed to assess the congruence between 26S rDNA and *matK-rbcL* sequence data. The ILD tests were conducted using the partition homogeneity

test on PAUP* following Mason-Gamer and Kellogg (1996). One thousand homogeneity test replicates were conducted using heuristic search with 100 random taxon additions, and TBR branch-swapping for each homogeneity replicate. Because an initial test suggested incongruence between the two data sets, further ILD tests for individual clades or excluding some clades from the matrix were conducted to identify lineages responsible for the incongruence. Both parsimony and ML analyses for combined data were conducted as described above.

RESULTS

Sequence data - The 26S rDNA sequences generated for the 68 species of Cornales and outgroups varied from 3340 to 3390 bp in length. The aligned matrix contained a total of 3430 bp, with small gaps between outgroups and Cornales taxa. Two additional insertions (bases 2096-2106 and 3251-3257) were detected in *Petalony*x. Sequences for most species were complete except those for *Decumaria* and one species of *Grubbia* (*G. rosmarinifolia*), which were missing approximately half of the 3´end. The sequences for *Hydrostachys angustisecta*-HS-4, *H. insignis* and *H. imbricata* were also incomplete with a portion of the 5´end missing. Including or excluding these taxa with incomplete sequences in the analyses did not affect the placements of remaining taxa in the resulting trees, thus we included them in the analyses for a broader sampling. Among the 68 sequences of Cornales and outgroups, 1048 of the 3430 sites are variable (30.56 %) and 557 sites (16.24%) are parsimony informative.

Twelve expansion segments (ES) were identified in the 26S rDNA sequence data matrix including outgroups. The expansion segments span a total of 1052 bp, of which 580 sites (55.13%) are variable and 393 sites (37.36%) are phylogenetically informative. These values are approximately 3-5 times higher than from core conserved (CC) regions, which contain 2378 bp, of which 468 sites (19.68%) are variable and only 164 sites (6.90%) are phylogenetically informative.

Among the 42 sequences of the combined data set (26S rDNA, *matK*, and *rbcL*), 2022 of the 6348 sites (31.85%) are variable and 1168 sites (18.40%) are phylogenetically informative. Among the 1168 phylogenetically informative sites, there are 466 from 26S rDNA, 454 from *matK*, and 248 from *rbcL*.

***Phylogenetic relationships based on 26S rDNA sequences*-**Parsimony analysis of 26S rDNA sequences alone found 47 most parsimonious trees of 2947 steps (Fig. 1). Eight major clades (supported by bootstrap support values of over 65%) were identified in all parsimonious trees: (1) *Cornus*; (2) *Alangium*; (3) nyssoids (*Nyssa*, *Davidia*, and *Camptotheca*); (4) mastixioids (*Diplopanax* and *Mastixia*); (5) *Curtisia* - *Grubbia*; (6) Loasaceae; (7) Hydrangeaceae; and (8) *Hydrostachys* (Fig.1). The relationships among these major clades suggested in the strict consensus tree are shown in Fig.1. None of the nodes connecting the major clades is supported by bootstrap analysis values of greater than 50% (Figs. 1 and 2). However, the differences among the 47 trees mostly involved only arrangements within Hydrangeaceae and among outgroup taxa. Compared to previous *matK*-*rbcL* based phylogeny, the strongly supported *Cornus*-*Alangium* clade is interrupted by *Hydrostachys*, which is placed as the sister of *Cornus* in the 26S rDNA trees (9% bootstrap

value, Figs. 1 and 2); the monophyly of Hydrangeaceae-Loasaceae is also contradicted by the 26S rDNA strict consensus trees.

ModelTest indicated GTR + I + Γ is the best fit model for the 26S rDNA sequence data. This GTR + I + Γ model incorporates both unequal base frequencies and different rates for all six substitutions and allows for among-site variation of substitution rates. A single best tree was found from the ML analysis using the GTR + I + Γ model and parameter values estimated from the model test. The same eight major clades as those found in the parsimony analysis were identified in the ML tree (Fig. 3), but the arrangements among these clades were different between the parsimony and ML trees. The monophyly of *Cornus-Alangium* was recovered, although without high bootstrap support (28%). The placement of *Hydrostachys* is dramatically different between the parsimony and ML trees. It is placed as the sister of *Cornus* in the parsimony analysis, whereas in the ML analysis it is placed as the sister of Loasaceae (Figs. 1 and 3). In both cases, this genus is monophyletic and connected by a long branch.

Cornus forms a monophyletic group, with four subclades: *C. canadensis*-*C. suecica*-*C. unalaschkensis* (the dwarf dogwoods); *C. mas*-*C. officinalis*-*C. sessilis* (the cornelian cherries), *C. florida*-*C. kousa*-*C. disciflora* (the big-bracted dogwoods), and *C. oblonga*-*C. racemosa*-*C. controversa*-*C. walteri* (the blue or white fruited group). Species of *Mastixia* also form a strongly supported monophyletic group in the mastixioids clade (BS = 100%), which is sister to *Diplopanax* with high bootstrap support (83% and 97%). *Nyssa* is monophyletic (98%, 100%), and *Camptotheca* and *Davidia* are sisters (57%, 60%) in the nyssoids clade. *Petalonyx* is sister to *Mentzelia* among the three sampled genera of Loasaceae. In Hydrangeaceae, subclade Hydrangeeae, consisting of *Platycrater*, *Decumaria*,

Pileostegia, *Schizophragma*, *Hydrangea*, *Broussaisia*, and *Cardiandra*, and subclade Philadelpheae, consisting of *Deutzia*, *Fendlerella*, and *Philadelphus*, were recognized and well-supported. The monophyly of Jamesioideae, however, was not supported in either parsimony or ML trees (Figs. 1-3). Within Hydrostachyaceae, three species from Madagascar (*H. angustisecta*, *H. imbricata,* and *H. multifida*) form a basal clade, are relative to the three Malawi species (*H. polymorpha, H. insignis,* and *H. augustisecta*) and one unidentified species from Madagascar (Figs. 1-3).

The analysis of ES regions using parsimony generated trees with topologies similar to those derived from the entire sequences (trees not shown). For example, the same eight major clades were similarly identified in the ES trees, and *Hydrostachys* was placed within Cornales. However, the ES trees have less resolution within and among major clades and lower bootstrap support for major clades than trees inferred from the entire sequences. The analysis of CC regions alone produced over 10000 trees without finishing searching, showing unexpected relationships within Cornales in the strict consensus tree, such as the collapse of strongly supported clades, including the *Cornus* clade (*C. volkensii* was separated from the other *Cornus* species, and placed in the outgroup), the nyssoids clade, and the Loasaceae clade (*Mentzelia* was placed as the most basal lineage of Cornales).

Incongruence test-The phylogenetic trees of Cornales inferred from 26S rDNA sequences were substantially different from those based on *matK* and *rbcL* sequences regarding the relationships among the major clades and within Hydrangeaceae (Xiang et al., 2002; also compare Figs. 3 and 4). Although the discrepancy mainly involved deep nodes that are mostly weakly supported in both cpDNA trees and 26S rDNA trees, we performed ILD tests to evaluate the congruence of the two data sets. The results indicated significant

incongruence between the *matK-rbcL* and 26S rDNA sequence data (P = 0.001). Subsequent successive ILD tests excluding individual major lineage one at a time were further conducted to locate the problematic lineages. Results revealed that much of the incongruence was attributed to a single group, Hydrangeaceae. The P -value of ILD tests increased (P = 0.003) only when Hydrangeaceae and outgroups were excluded (Table 3). Further, ILD tests were performed for each major lineage and results also showed that Hydrangeaceae, in particular the subclade Hydrangeeae, is the only ingroup showing significant disagreement between the two data sets (Table 3).

***Phylogenetic relationships based on combined 26S rDNA, matK* and *rbcL sequences*-** Considering that the topological discrepancy between 26S rDNA and *matK-rbcL* trees mainly involved weakly supported nodes, and potentially only a single lineage exhibits conflicts between the two data sets, we performed analyses of the combined 26S rDNA and *matK-rbcL* sequences. A single most parsimonious tree was found from the parsimony analysis of combined data [tree length 4735, CI (consistency index) = 0.570, RI (retention index) = 0.547] (Fig. 5). The tree shows the monophyly of *Cornus-Alangium*, Loasaceae-Hydrangeaceae (weakly supported), nyssoids-mastixioids, and *Grubbia-Curtisia*. *Hydrostachys* groups with outgroups, sister to the remainder of Cornales (Fig. 5). Model test similarly suggested that the GTR + I + Γ model best fits the combined data. The ML analysis using this model resulted in a single tree (Fig. 6) with topology showing the same eight major clades and relationships within and among the clades similar to those in the *matK-rbcL* tree (Fig. 4). *Hydrostachys* was placed at the base of Cornales with low bootstrap supports (Fig. 6). However, bootstrap and CI values increased significantly for most clades in the combined 26S rDNA-*matK-rbcL* trees (compare Figs. 2 and 3 with 5 and 6).

DISCUSSION

Differential phylogenetic utility of the ES and CC regions of 26S rDNA in Cornales- The large-subunit of rDNA is structured as a mosaic of core conserved and variable domains (as defined as expansion segments by Clark et al., 1984). The "core" conserved segments have primary and secondary structures conserved in prokaryotes and eukaryotes (Hancock and Dover, 1990). The expansion segments are responsible for the difference in size between eukaryotic and prokaryotic rDNA, and they evolve much faster than conserved core regions. Despite their rapid evolution, expansion segments still contain a conserved secondary structure and show nucleotide composition constraints in some species (Hassouna et al., 1984; Gorab et al., 1995). Comparative analyses confirmed that the base substitution rates in the expansion segments of 26S rDNA are lower than those observed in nuclear noncoding regions or neutral bases (Larson and Wilson, 1989). Compared to animals, the considerable length mutation of the expansion segments observed in animal rDNA is not found in angiosperms (Kolosha and Fodor, 1990), and consequently, the expansion segments may be more alignable among angiosperms and the point mutations in the sequence may be phylogenetically informative at different taxonomic levels in plants (Kuzoff at al., 1998). Because of the relatively high rate of substitutions in the expansion segments, Larson (1991) suggested that the expansion regions should be excluded from phylogenetic analyses of taxa with a common ancestor older than 200 MYA (million years ago) due to possible saturation of substitutions. The 26S rDNA in Cornales contains 12 expansion segments, which evolve about 3-5 times as fast as the conserved core regions. This ratio is much lower than that

observed for other angiosperms (6.4 to 10.2 times, Kuzoff et al., 1998). The analyses of separate partitions of ES and CC regions using parsimony suggested that in Cornales most of the phylogenetic signals of 26S rDNA are from the ES regions and the CC regions alone are not sufficient to resolve meaningful relationships in the group, although it might be at higher taxonomic levels.

Discrepancies between parsimony and maximum likelihood analyses-It is well recognized that one potential problem of parsimony analysis is the inconsistency of the method if substitution rates are high and unequal among lineages (Felsenstein, 1978; Swofford et al., 1996; Swofford et al., 2001). In this case, unrelated taxa with high rates (shown as long branches in the data matrix) will be likely attracted to each other in a simple parsimony analysis. ML analysis implementing appropriate substitution model(s) is supposed to be able to largely overcome this long branch problem (Felsenstein, 1981; Swofford et al., 1996). Our analyses of 26S rDNA sequences and combined 26S rDNA-*matK-rbcL* sequences using parsimony and ML methods suggested different placement for the long-branched *Hydrostachys*. Parsimony analysis of 26S rDNA sequence data placed *Hydrostachys* in the *Cornus-Alangium* clade, whereas the ML analysis of 26S rDNA data placed it with the Loasaceae (Figs. 1-3). Analyses of the combined 26S rDNA-*matK-rbcL* using parsimony grouped *Hydrostachys* with outgroups, whereas ML analysis of the combined data grouped it with Cornales and placed it as the sister to the remainder of the Cornales clade. The placements of *Hydrostachys* in the 26S rDNA and combined data are both weakly supported. Additional discrepancies of relationships among major lineages between parsimony and ML analyses were also found in our analyses. For example, the sister relationship between *Cornus* and *Alangium* identified in the ML analysis was

congruent with all previous chloroplast data analyses and supported by morphological characters (Eyde, 1988). Nevertheless, this relationship was broken off by *Hydrostachys* in 26S rDNA parsimony analysis. The monophyly of traditional cornalean taxa including *Alangium*, *Cornus*, nyssoids, and mastixioids recognized in the ML trees (Fig. 3) are in agreement with morphology and previous chloroplast data analysis, but the monophyly of these taxa was not identified in the parsimony trees (Figs. 1, 2, and 5). However, it must be noted that these relationships showing discrepancies are generally not strongly supported in either parsimony or ML trees.

Placements of Hydrostachyaceae and the long branch-The systematic affinity of the African aquatic family Hydrostachyaceae (consisting of only *Hydrostachys* with 22-25 species) has long been controversial. It has been placed near Podostemaceae (Bentham and Hooker, 1880), or as the distinct order Hydrostachyales allied with Lamiales and Scrophulariales (Takhtajan, 1969, 1980, 1997; Dahlgren, 1980, 1983; Leins and Erbar, 1988, 1990; Dahlgren, 1989; Wagenitz, 1992) and in Bruniales (Thorne, 1968, 1983, 1992, 2000) and Callitrichales (Cronquist, 1981) based on the features in morphology. The family was first linked to Cornales in the *rbcL* sequence analysis of Loasaceae by Hempel et al. (1995). More recent molecular data analyses (Xiang, 1999; Albach et al., 2001a, b; Xiang et al., 2002) and evidence from phytochemistry (Rønsted et al., 2002) further supported the placement of this family in Cornales. A majority of previous phylogenetic analyses suggested a position of *Hydrostachys* within Hydrangeaceae, with low bootstrap support (e.g., Hempel et al., 1995; Xiang, 1999; Xiang et al., 2002). A few possible synapomorphies of Hydrangeaceae and *Hydrostachys* were identified by previous authors, such as two or more free styles, capsules, and numerous, anatropous ovules per locule (see Xiang, 1999;

Albach et al., 2001a). ML analyses of 26S rDNA sequence data and combined 26S rDNA-*matK*-*rbcL* sequence data, as well as parsimony analysis of 26S rDNA data, all suggested that *Hydrostachys* is a member of Cornales but revealed new placements in Cornales different from those suggested in previous analyses. For example, the ML tree of 26S rDNA placed *Hydrostachys* as sister to Loasaceae (Fig. 3). The ML tree of the combined 26S rDNA-*matK*-*rbcL* shows that *Hydrostachys* is sister to the remainder of Cornales (Fig. 6). The placement of *Hydrostachys* with outgroups in the parsimony analysis of the combined data is likely a result of long-branch attraction, given that the branches leading to *Hydrostachys* and the outgroup clade are both long (Fig. 5).

As discussed above, long-branch attraction is a concern in phylogenetic analyses using a parsimony approach (Felsenstein, 1978; Swofford et al., 2001). Both simulation (e.g., Hillis and Huelsenbeck, 1993; Huelsenbeck, 1995; Yang, 1996; Siddall, 1998; Pol and Siddall, 2001) and empirical studies (e.g., Omilian and Taylor, 2001; Litvaitis, 2002) have shown that long-branch attraction can result in wrong phylogenies when using a parsimony method. One recommended solution to long-branch attraction is to increase the sampling of long-branched taxa to decrease the branch length. In our 26S rDNA analysis, seven species of *Hydrostachys* were sampled with the attempt to reduce the long branch of the genus revealed in previous various cpDNA analyses (Xiang, 1999; Xiang et al., 2002). With increasing sampling, the branches leading to *Hydrostachys* in the parsimony and ML 26S rDNA trees were significantly reduced in length compared to those in the cpDNA trees with only one or two species sampled (Xiang, 1999; Albach et al., 2001a,b; Xiang et al., 2002). In the 26S rDNA trees, the branch of *Hydrostachys* is not much longer than the outgroup branches (Figs. 2 and 3), and only about twice as long as the longest ingroup branches (Figs. 2 and 3).

In all phylogenetic analyses of cpDNA sequence data sampling a single or two species (e.g., Xiang, 1999; Xiang et al., 2002; Figs. 4-6), the branches of *Hydrostachys* were much longer, sometimes several times longer, than the longest branches of the ingroups, and much longer than the longest outgroup branches. These results demonstrated that increasing sampling of the long-branched group indeed substantially decreased the branch length.

Many studies have suggested that organisms that are highly modified may morphologically have accelerated rates of molecular evolution (Nickrent and Starr, 1994; DePamphilis et al., 1997; Les et al., 1997; Mallat and Sullivan, 1998; Soltis et al., 1999, 2000; Chase et al., 2000; Albach et al., 2001a). *Hydrostachys*, due to its aquatic habit, is morphologically highly divergent from the remaining cornalean taxa (e.g., pinnate compound leaves, tuber-like rhizomes, and dense spike inflorescence). Its long branches revealed in previous analyses could be viewed as evidence of its elevated rates of molecular evolution in the genus. However, long branches could be simply a result of the incomplete sampling from the genus, as increasing sampling substantially reduced the branch length in our 26S rDNA analysis. However, this sampling effect is less clear when examining the trees from combined nuclear and cpDNA sequences (Figs. 5 and 6). Based on the combined 26S rDNA-*matK*-*rbcL* sequence data, the separation of *Hydrostachys* from the rest of Cornales might have occurred very early, before the origin of all other cornalean major lineages (Fig. 6).

Relationships of Grubbia and Curtisia-Grubbiaceae, another monogeneric family of Cornales from southern Africa, in addition to Curtisiaceae and Hydrostachyaceae, represents another family difficult to place in the classification of flowering plants. Both separate and combined data analyses in the present study suggested that *Grubbia* and *Curtisia* are sisters,

in agreement with the previous finding from the *matK-rbcL* data (Xiang et al., 2002). The sister relationship between *Grubbia* and *Curtisia* is supported by high bootstrap values in all analyses. Unlike *Hydrostachys*, which shows no apparent morphological similarities with other cornalean taxa, *Grubbia* and *Curtisia* share several morphological features that are common in the Cornales (see Xiang, 1999). Therefore, the finding of a close relationship between the two genera both endemic to southern Africa is not a surprise. The circumscription of Grubbiaceae including both *Grubbia* and *Curtisia* as proposed by Xiang et al. (2002) is strongly supported. Relationships of *Grubbia-Curtisia* to other cornalean taxa are not clearly resolved.

Monophyly of nyssoids, mastixioids, Cornus, and Alangium-The monophyly of nyssoids, mastixioids, *Cornus*, and *Alangium* was suggested in ML analysis of 26S rDNA data (Fig. 3). This clade is also supported by a few nonmolecular characters (e.g., fleshy drupaceous fruit with germination valves on fruit stones, H-shaped thinning in pollen aperture, and the lack of central bundles in gynoecial vasculature), and largely corresponds to the Cornaceae of Eyde (1988). However, given the low bootstrap support for the clade (Fig. 3), it is better to maintain the nyssoids and mastixioids as separate families as discussed in Xiang et al. (2002).

The monophyly of the nyssoids, mastixioids, and *Cornus–Alangium* subclades are strongly supported in the combined 26S rDNA-*matK-rbcL* data analyses. The sister relationship between *Cornus* and *Alangium* has been also recognized in previous molecular studies (Xiang, 1999; Xiang et al., 1993, 1998, 2002) and is also supported by some morphological and embryological characters (e.g., unitegmic and crassinucellate ovules; degeneration of nucleus followed by the differentiation of an integumentary tapetum; single-

celled archesporium; see Chopra and Kaur, 1965; Eyde, 1968, 1988). Based on this evidence, Xiang et al. (2002) proposed a Cornaceae consisting of *Cornus* and *Alangium* following Soltis et al. (2000). Because *Alangium* has long been recognized as a monogeneric family and the name Alangiaceae has been widely used, we proposed here to separate *Cornus* and *Alangium* in Cornaceae and Alangiaceae, respectively.

The relationships within the nyssoids vary between separate data partitions and combined data. In analysis of 26S rDNA sequences, *Camptotheca* is sister to *Davidia* (57%, 60%; Figs. 2 and 3), whereas in analyses of combined 26S rDNA-*matK*-*rbcL* sequence data, *Nyssa* is strongly supported to be the sister of *Camptotheca* (BS = 83%), and the two, in turn, are sister to *Davidia* (Figs. 5-6). These relationships were also found in earlier and present analyses of *matK* and *rbcL* sequences (Xiang et al., 1998, 2002). A closer relationship of *Camptotheca* to *Nyssa* is also supported by some nonmolecular data (e.g., the structure of the fruits and the inflorescences: Eyde, 1963, 1967; wood anatomy: Titman, 1949; palynology - Eramian, 1971; Eyde and Barghoorn, 1963; and fatty acids: Bate-Smith et al., 1975; Hohn and Meinschein, 1976).

The sister relationship between *Mastixia* and *Diplopanax* (Fig. 6) was first recovered in the combined *rbcL*-*matK* sequence analysis of Xiang et al. (2002), and again recovered in the present study with high bootstrap support in all analyses. The close relationship between *Mastixia* and *Diplopanax* was earlier recognized by Eyde and Xiang (1990) and further supported by Zhu and Xiang (1999) via studies of fruit, leaf and floral anatomic structures. Both genera produce flowers with hooked petals that are arranged in paniculate inflorescences, fruits have a bony stone with an intrusive germination valve lacking a longitudinal septum, and a one-seeded chamber.

Phylogenetic relationships in Hydrangeaceae and Loasaceae-Two strongly supported monophyletic groups, which correspond to the two tribes, Hydrangeeae and Philadelpheae, were recognized in Hydrangeaceae in both separate and combined analyses. The monophyly of Jamesioideae was not recognized in the 26S rDNA sequence analyses, but was in the tree based on combined data (Figs. 1-6). The relationships within Hydrangeeae suggested by 26S rDNA and *matK-rbcL* were different (Figs. 1-4). The combined 26S rDNA-*matK-rbcL* data agreed with the *matK-rbcL* data in placing *Pileostegia* + *Decumaria* as the sister of *Schizophragma* with high bootstrap support (Figs. 4, 5, and 6). The close relationships among the genera are also supported by morphological data (Hufford, 1992, 1997) and recovered in previous phylogenetic analyses (Soltis et al., 1995; Hufford et al., 2001; Xiang et al., 2002). Morphological data suggested that *Platycrater* was outside of the *Hydrangea* clade (including genera of *Hydrangea*, *Pileostegia*, *Decumaria*, *Broussaisia*, and *Schizophragma* in this study), a clade supported by a synapomorphic character of diplostemony (Hufford, 1997). However, all molecular analyses (Soltis et al., 1995; Xiang, 1999; Hufford et al., 2001; and the present study) placed *Platycrater* within the *Hydrangea* clade, suggesting that diplostemony might have been lost in *Platycrater*, as previously hypothesized by Hofford et al. (2001). The relative relationships among *Hydrangea*, *Broussaisia*, and *Cardiandra* are different in 26S rDNA and combined 26S rDNA-*matK-rbcL* trees all with strong bootstrap supports (Figs. 1-3, 5, and 6). However, these relationships were not revealed in previous phylogenetic analyses with a more thorough sampling of genera of Hydrangeaceae (Soltis et al., 1995; Hufford et al., 2001; Xiang et al., 2002). In those analyses with a complete sampling of genera in the family, *Cardiandra* and *Deinanthe* were recognized as sisters and placed at the base within the Hydrangeeae clade

(Hufford et al., 2001; Xiang et al., 2002). Therefore, the sister relationships among these three taxa revealed in the present study is likely a result of incomplete sampling.

Only four species of Loasaceae representing two of the three subfamilies (Gronovioideae and Mentzelioideae), were sampled in this study, thus relationships within Loasaceae cannot be appropriately addressed with confidence. However, the two genera, *Eucnide* and *Mentzelia*, from subfamily Mentzelioideae, do form a monophyletic group in the combined data analyses (bootstrap value 100%). The two are, in turn, sister to *Petalonyx* (from subfamily Gronovioideae). These relationships are also congruent with earlier studies using *rbcL* sequence data (Hempel et al., 1995) and our *matK-rbcL* sequence analysis (Fig. 3). However, 26S rDNA sequence data alone placed *Mentzelia* sister to *Petalonyx*, agreeing with the *matK* and ITS sequence data (Moody et al., 2001). *Eucnide* and *Mentzelia* share many morphological characters in floral structures (e.g. polystemonous, multicarpellate, multiovulate, and dehiscent fruits). However, some possible morphological synapomorphies (e.g. the absence of the petal-stamen plate in *Mentzelia* and Gronovioideae) may unite *Mentzelia* and Gronovioideae (including *Petalonyx*). This discrepancy may be due to either inadequate sampling of Loasaceae in different studies and/or different phylogenetic signals between data sets which needs further investigation.

Conclusion-Phylogenetic analyses of nuclear DNA sequence data and combined nuclear and chloroplast DNA sequence data further support a Cornales consisting of *Cornus*, *Alangium*, nyssoids, mastixioids, Hydrangeaceae, Loasaceae, Grubbiaceae *(Grubbia-Curtisia*), and Hydrostachyaceae (*Hydrostachys*). Four most-inclusive major clades in Cornales (*Cornus-Alangium*, nyssoids-mastixioids, Hydrangeaceae-Loasaceae, and *Grubbia-Curtisia*) identified in previous *matK-rbcL* sequence analyses (Xiang et al., 1998; Xiang,

1999; Xiang et al., 2002), were also recovered in analyses of the combined nuclear and chloroplast DNA sequence data in the present study. The combined 26S rDNA-*matK*-*rbcL* sequence data suggested that Hydrostachyaceae probably branched early from the remainder of Cornales. Relationships among major lineages of Cornales are weakly supported by bootstrap analyses, similar to previous studies. This uncertainty of relationships among major lineages of Cornales, despite rigorous analyses of a large number of characters, may reflect an early rapid radiation of the Cornales clade. The present study supports the classification within Cornales proposed in Xiang et al. (2002): a Cornaceae of *Cornus*-*Alangium*, a Nyssaceae consisting of *Nyssa*, *Davidia*, and *Camptotheca*, a Mastixiaceae consisting of *Mastixia* and *Diplopanax*, a Grubbiaceae including *Curtisia* and *Grubbia*, Hydrangeaceae, Loasaceae, and Hydrostachyaceae. Given that Alangiaceae has long been widely used and there are also many morphological differences between *Alangium* and *Cornus* (e.g., leaf arrangement nearly always opposite for *Cornus*, alternate for *Alangium*; stamens isomerous with the perianth in *Cornus*, but mostly 2-4 times of perianth parts in *Alangium*; and inflorescence mostly terminal in *Cornus*, but mostly lateral in *Alangium*), it is more desirable to maintain *Cornus* and *Alangium* as two distinct families. Our study also indicated the following: (1) increased sampling of *Hydrostachys* species reduced its long branch length substantially; (2) combining data significantly increased bootstrap support and CI value; (3) major discrepancies between parsimony and maximum likelihood analyses were found regarding the placement of long-branched taxa (e.g., *Hydrostachys*).

REFERENCES CITED

ALBACH, D. C., D. E. SOLTIS, M. W. CHASE, AND P. S. SOLTIS. 2001a. Phylogenetic placement of the enigmatic angiosperm *Hydrostachys*. *Taxon* 50: 781-805.

ALBACH, D. C., P. S. SOLTIS, D. E. SOLTIS, AND R. G. OLMSTEAD. 2001b. Phylogenetic analysis of the asterids based on sequences of four sequences. *Annals of the Missouri Botanical Garden* 88: 163-210.

APG. 1998. An ordinal classification for the families of flowering plants. *Annals of the Missouri Botanical Garden* 85: 531-553.

ASHWORTH, V. E. T. M. 2000. Phylogenetic relationship in Phoradendreae (Viscaceae) inferred from three regions of the nuclear ribosomal cistron. I. Major lineages and paraphyly of Phoradendron. *Systematic Botany* 25: 349-370.

BATE-SMITH, E. C., I. K. FERGUSON, K. HUTSON, S. R. JENSEN, B. J. NIELSEN, AND T. SWAIN. 1975. Phytochemical interrelationships in the Cornaceae. *Biochemical Systematics and Ecology* 3: 79-89.

BENTHAM, G., AND J. D. HOOKER. 1880. Genera plantarum, vol. 3. London, England, UK.

CHANDERBALI, A. S., H. VAN DER WERFF, AND S. S. RENNER. 2001. Phylogeny and historical biogeography of Lauraceae: evidence from the chloroplast and nuclear genomes. *Annals of the Missouri Botanical Garden* 81: 104-134.

CHASE, M., M. F. FAY, AND V. SAVOLAINEN. 2000. Higher-level classification in the angiosperms: new insights from the perspective of DNA sequence data. *Taxon* 49: 685-704.

CHASE, M., ET AL. 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbcL*. *Annals of the Missouri Botanical Garden* 80: 528-580.

CHOPRA, R. N., AND H. KAUR. 1965. Some aspects of the embryology of *Cornus*. *Phytomorphology* 15: 353-359.

CLARK, G. B., B. W. TAGUE, V. C. WARE, AND S. A. GERBI. 1984. *Xenopus Laevis* 28S ribosomal RNA: a secondary structure and its evolutionary and functional implications. *Nucleic Acids Research* 12: 6197-6220.

CRONQUIST, A. 1981. An integrated system of classification of flowering plants. Columbia University Press, New York, New York, USA.

CULLINGS, K. W. 1992. Design and testing of a plant-specific PCR primer for ecological and evolutionary studies. *Molecular Ecology* 1: 233-240.

DAHLGREN, G. 1989. The last dahlgrenogram, a system of classification of the dicotyledons. *In* K. Tan [ed.], Plant taxonomy, phytogeography and related subjects: the Davis and Hedge festschrift, 249-260. Edinburgh University Press, Edinburgh, Ireland.

DAHLGREN, R. M. T. 1980. A revised system of classification of the angiosperm. *Botanical Journal of the Linnaeus Society* 80: 91-124.

DAHLGREN, R. M. T. 1983. General aspects of angiosperm evolution and macrosystematics. *Nordic Journal of Botany* 3: 119-149.

DEPAMPHILIS, C., N. D. YOUNG, AND A. D. WOLFE. 1997. Evolution of plastid gene *rps*2 in a lineage of hemiparasitic and holoparasitic plants: many losses of photosynthesis and complex patterns of rate variation. *Proceedings of National Academy of Sciences, USA* 94: 7367-7372.

DOVER, G. A., AND R. B. FLAVELL. 1984. Molecular coevolution: DNA divergence and the maintenance of function. *Cell* 38: 622-623.

DOWNIE, S. R., AND J. D. PALMER. 1992. Restriction site mapping of the chloroplast DNA inverted repeat: a molecular phylogeny of the Asteridae. *Annals of Missouri Botanical Garden* 79: 266-283.

ERAMIAN, E. N. 1971. Palynological data on the systematics and phylogeny of Cornaceae Dumort. and related families. *In* L. A. Kuprianova and M. S. Yakovlev [eds.], Pollen morphology of Cucurbiatceae, Thymelaeaceae, Cornaceae, 235-273. Leningrad, Russia.

EYDE, R. H. 1963. Morphological and paleobotanical studies of the Nyssaceae, I. A survey of the modern species and their fruits. *Journal of the Arnold Arboretum* 44: 1-59.

EYDE, R. H. 1967. The peculiar gynoecial vasculature of Cornaceae and its systematic significance. *Phytomorphology* 17: 172-182.

EYDE, R. H. 1968. Flower, fruits and phylogeny of Alangiaceae. *Journal of the Arnold Arboretum* 49: 167-192.

EYDE, R. H. 1988. Comprehending *Cornus*: puzzles and progress in the systematics of the dogwoods. *Botanical Review* 3: 233-351.

EYDE, R. H., AND E. S. BARGHOORN. 1963. Morphological and paleobotanical studies of the Nyssaceae, II. The fossil record. *Journal of the Arnold Arboretum* 44: 328-376.

EYDE, R. H., AND Q.-Y. Xiang. 1990. Fossil mastixioid (Cornaceae) alive in eastern Asia. *American Journal of Botany* 77: 689-692.

FAN, C., AND Q.-Y. XIANG. 2001. Phylogenetic relationships within *Cornus* (Cornaceae) based on 26S rDNA sequences. *American Journal of Botany* 88: 1131-1138.

FARRIS, J. S., M. KALLERSJO, A. G. KLUGE, AND C. BULT. 1994. Testing significance if incongruence. *Cladistics* 10: 315-319.

FELSENSTEIN, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* 27: 401-410.

FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequence: a maximum likelihood approach. *Journal of Molecular Evolution* 17: 368-376.

FELSENSTEIN, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39: 783-791.

FISHBEIN, M, C. HIBSCH-JETTER, D. E. SOLTIS, AND L. HUFFORD. 2001. Phylogeny of Saxifragales (Angiosperms, Eudicots): Analysis of a rapid, ancient radiation. *Systematic Biology* 50: 817-847.

FLAVELL, R. B. 1986. Structure and control of expression of ribosomal RNA genes. *Oxford Survey of Plant Molecular Cell Biology* 3: 252-274.

GORAB, E., M. G. DE LACOBA, AND L. M. BOTELLA. 1995. Structural constraints in expansion segments from a midge 26S rDNA. *Journal of Molecular Evolution* 41: 1016-1021.

HANCOCK, J. M., AND G. A. DOVER. 1990. 'Compensatory slippage' in the evolution of ribosomal RNA gene. *Nucleic Acids Research* 18: 5949-5954.

HARMS, H. 1898. Cornaceae. *In* A. Engler and K. Prantl [eds.], Die natürlichen Pflanzenfamilien, III.8, 250-270. Wilhelm Engelmann, Leipzig, Germany.

HASSOUNA, N., B. MICHOT, AND J. BACHELLERIE. 1984. The complete nucleotide sequence of mouse 28S rRNA gene: implications for the process of size increase of the large subunit rRNA in higher eukaryotes. *Nucleic Acids Research* 12: 3563-3574.

HEMPEL, A. L., P. A. REEVES, R. G. OLMSTEAD, AND R. K. JANSEN. 1995. Implications of *rbcL* sequence data for higher order relationships of the Loasaceae and the anomalous aquatic plant *Hydrostachys* (Hydrostachyaceae). *Plant Systematics and Evolution* 194: 25-37.

HILLIS , D. M., AND J. P. HUELSENBECK. 1993. Success of phylogenetic methods in the four-taxon case. *Systematic Biology* 42: 247-264.

HOHN, M. E., AND W. G. MEINSCHEIN. 1976. Seed oil fatty acids: Evolutionary significance in the Nyssaceae and Cornaceae. *Biochemical Systematics and Ecology* 4: 193-199.

HUELSENBECK, J. P. 1995. Performance of phylogenetic methods in simulation. *Systematic Biology* 44: 17-48.

HUFFORD, L. D. 1992. Rosidae and their relationship to other nonmagnoliid dicotyledons: a phylogenetic analysis using morphological and chemical data. *Annals of Missouri Botanical Garden* 79: 219-248.

HUFFORD, L. D. 1997. A phylogenetic analysis of Hydrangeaceae using morphological data. *International Journal of Plant Sciences* 158: 652-672.

HUFFORD, L. D, M. L. MOODY, AND D. E. SOLTIS. 2001. A phylogenetic analysis of Hydrangeaceae based on sequences of the plastid gene *matK* and their combination with *rbcL* and morphological data. *International Journal of Plant Sciences* 162: 835-846.

HUTCHINSON, J. 1967. The genera of flowering plants. Clarendon Press, Oxford, UK.

KOLOSHA, V. O., AND I. FODOR. 1990. Nucleotide sequence of *Citrus limon* 26S rDNA gene and secondary structure model of its RNA. *Plant Molecular Biology* 14: 147-161.

KUZOFF, R. K., J. A. SWEERE, D. E. SOLTIS, P. S. SOLTIS, AND E. A. ZIMMER. 1998. The phylogenetic potential of entire 26S rDNA sequences in plant. *Molecular Biology and Evolution* 15: 251-263.

LARSON, A. 1991. Evolutionary analysis of length variable sequences: divergence domains of ribosomal RNA. *In* M. Miyamoto and J. Cracraft [eds.], Phylogenetic analysis of DNA sequences, 221-247. Oxford University Press, New York, New York, USA.

LARSON, A., AND A. C. WILSON. 1989. Patterns of ribosomal RNA evolution in salamanders. *Molecular Biology and Evolution* 6: 131-154.

LES, D. H., C. T. PHILBRICK, AND A. R. NOVELO. 1997. The phylogenetic position of river-weeds (Podostemaceae): insights from *rbcL* sequence data. *Aquatic Botany* 57: 5-27.

LEINS, V. P., AND C. ERBAR. 1988. Some remarks on flower development and systematic position of the water plants *Callitriche*, *Hippuris* and *Hydrostachys*. *Beitrage zur Biologie der Pflanzen* 63: 157-178.

LEINS, V. P., AND C. ERBAR. 1990. The possible relationship of Hydrostachyaceae based on comparative ontogenetical flower studies. *Proceedings of the twelfth plenary meeting of AETFAT Mitteilungen aus dem Institut für Allgemeine Botanik, Hamburg* 23b: 723-729.

LITVAITIS, M. K. 2002. A molecular test of cyanobacterial phylogeny: inference from constraint analyses. *Hydrobiologia* 468: 135-145.

MALLATT, J., AND J. SULLIVAN. 1998. 28S and 18S rDNA sequences support the monophyly of lampreys and hagfishes. *Molecular Biology and Evolution* 15: 1706-1718.

MASON-GAMER, R. J., AND E. A. KELLOGG. 1996. Testing for phylogenetic conflict among molecular data sets in the tribe Triticeae (Granineae). *Systematic Biology* 45: 524-545.

MICKEVICH, M. F., AND J. S. FARRIS. 1981. The implications of congruence in *Menidia*. *Systematic Zoology* 30: 351-370.

MISHLER, B. D., L. A. LEWIS, M. A. BUCHHEIM, K. S. RENZAGLIA, D. J. GARBARY, C. F. DELWIICHE, F. W. ZECHMAN, T. S. KANTZ, AND R. L. CHAPMAN. 1994. Phylogenetic relationships of the "green algae" and "bryophytes." *Annals of the Missouri Botanical Garden* 81: 451-483.

MOODY, M. L., L. HUFFORD, D. E. SOLTIS, AND P. S. SOLTIS. 2001. Phylogenetic relationships of Loasaceae subfamily Gronovioideae inferred from *matK* and ITS sequence data. *American Journal of Botany* 88: 326-336.

MORGAN, D. R., AND D. E. SOLTIS. 1993. Phylogenetic relationships among members of Saxifragaceae sensu lato based on *rbcL* sequence data. *Annals of the Missouri Botanical Garden* 80: 631-660.

MORTON, C. M., M. W. CHASE, K. A. KRON, AND S. M. SWENSEN. 1996. A molecular evolution of the monophyly of the order Ebenales based upon *rbcL* sequence data. *Systematic Botany* 21: 567-586.

NEYLAND, R. 2001. A phylogeny inferred from large ribosomal subunit (26S) rDNA sequences suggests that *Cuscuta* is a derived member of Convolvulaceae. *Brittonia* 53: 108-115.

NICKRENT, D. L., AND E. M. STARR. 1994. High rates of nucleotide substitution in nuclear small-subunit (18S) rDNA from holoparasitic flowering plants. *Journal of Molecular Evolution* 39: 62-70.

NICKRENT, D. L., A. BLARER, Y-L. QIU, D. E. SOLTIS, P. S. SOLTIS, AND M. ZANIS. 2002. Molecular data place Hydnoraceae with Aristolochiaceae. *American Journal of Botany* 89: 1809-1817.

OLMSTEAD, R. G., B. BREMER, K. M. SCOTT, AND J. D. PALMER. 1993. A parsimony analysis of the Asteridae sensu lato based on *rbcL* sequences. *Annals of Missouri Botanical Garden* 80: 700-722.

OLMSTEAD, R. G., K-J. KIM, R. K JANSEN, AND S. J. WAGSTAFF. 2000. The phylogeny of the Asteridae sensu lato based on chloroplast *ndhF* gene sequences. *Molecular Phylogenetics and Evolution* 16: 96-112.

OMILIAN, A. R., AND D. J. TAYLOR. 2001. Rate acceleration and long-branch attraction in a conserved gene of cryptic daphniid (Crustacea) species. *Molecular Biology and Evolution* 18: 2201-2212.

POL, D., AND M. E. SIDDALL. 2001. Bias in maximum likelihood and parsimony: a simulation approach to a 10-taxon case. *Cladistics* 17: 266-281.

POSADA, D., AND K. A. CRANDALL. 1998. Modeltest: testing the model of DNA substitution. *Bioinformatics* 14: 817-818.

RO, K-E., H-Y HAN, AND S. LEE. 1999. Phylogenetic contributions of partial 26S rDNA sequences to the tribe Helleboreae (Ranunculaceae). *Korean Journal of Biological Sciences* 3: 9-15.

RO, K-E., C. S. KEENER, AND B. A. MCPHERON. 1997. Molecular phylogenetic study of the Ranunculaceae: utility of the nuclear 26S ribosomal DNA in inferring intrafamilial relationships. *Molecular Phylogenetics and Evolution* 8: 117-127.

RØNSTED, N., H. STRANDGAARD, S. R. JENSEN, AND P. MØLGAARD. 2002. Clorogenic acid from three species of *Hydrostachys*. *Biochemical Systematics and Ecology* 30: 1105-1108.

SAVOLAINEN, V., M. W. CHASE, S. B. HOOT, C. M. MORTON, D. E. SOLTIS, C. BAYER, M. F. FAY, A. Y. DE BRUIJN, S. SULLIVAN, AND Y-L. QIU. 2000a. Phylogenetics of flowering plants based upon a combined analysis of plastid *atpB* and *rbcL* gene sequences. *Systematic Biology* 49: 306-362.

SAVOLAINEN, V., ET AL. 2000b. Phylogeny of the eudicots: a nearly complete familial analysis based on *rbcL* gene sequences. *Kew Bulletin* 55: 257-309.

SIDDALL, M. E. 1998. Success of parsimony in the four-taxon case: long-branch repulsion by likelihood in the Farris zone. *Cladistics* 14: 209-220.

SIMMONS, M. P., V. SAVOLAINEN, C. C. CLEVINGER, R.H. ARCHER, AND J. I. DAVIS. 2001. Phylogeny of the Celastraceae inferred from 26S nuclear ribosomal DNA, phytochrome B, *rbcL*, *atpB*, and morphology. *Molecular Phylogenetics and Evolution* 19: 353-366.

SOLTIS, D. E., AND P. S. SOLTIS. 1997. Phylogenetic relationships among Saxifragaceae sensu lato: a comparison of topologies based in 18S rDNA and *rbc*L sequences. *American Journal of Botany* 84:504-522.

SOLTIS, D. E., AND P. S. SOLTIS. 1998. Choosing an approach and an appropriate gene for phylogenetic analysis. *In* D. E. Soltis, P. E. Soltis, and J. J. Doyle [eds.], Molecular systematics of plants II, 1-42. Chapman and Hall, New York, New York, USA.

SOLTIS, D. E., R. K. KUZOFF, M. E. MORT, M. ZANIS, M. FISHBEIN, L. HUFFORD, J. KOONTZ, AND M. K. ARROYO. 2001. Elucidating deep-level phylogenetic relationships in Saxifragaceae using sequences for six chloroplastic and nuclear DNA regions. *Annals of the Missouri Botanical Garden* 88: 669-693.

SOLTIS, D. E., M. E. MORT, P. S. SOLTIS, C. HIBSCH-JETTER, E. A. ZIMMER, AND D. MORGAN. 1999. Phylogenetic relationships of the enigmatic angiosperm family Podostemaceae inferred from 18S rDNA and *rbcL* sequence data. *Molecular Phylogenetics and Evolution* 11: 261-272.

SOLTIS, D. E., ET AL. 2000. Angiosperm phylogeny inferred from 18S rDNA, *rbcL*, and *atpB* sequences. *Biological Journal of the Linnean Society* 133: 381-461.

SOLTIS, D. E., Q.-Y. XIANG, AND L. HUFFORD. 1995. Relationships and evolution of Hydrangeaceae based on *rbcL* sequence data. *American Journal of Botany* 82: 504-514.

STEFANOVIC, S., M. JAGER, J. DEUTSCH, J. BROUTIN, AND M. MASSELOT. 1998. Phylogenetic relationships of conifers inferred from partial 28S rDNA gene sequences. *American Journal of Botany* 85: 688-697.

SWOFFORD, D. L. 2002. PAUP: phylogenetic analysis using parsimony, version 4.0b10. Sinauer Associates, Sunderland, Massachusetts, USA.

SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, AND D. M. HILLIS. 1996. Phylogenetic inference. *In* D. H. Hillis, C. Moritz, and B. K. Mable [eds.], Molecular systematics, 407-514. Sinauer Associates, Sunderland, Massachusetts, USA.

SWOFFORD, D. L., P. J. WADDELL, J. P. HUELSENBECK, P. G. FOSTER, P. O. LEWIS, AND J. S. ROGERS. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Systematic Biology* 50: 525-539.

TAKHTAJAN, A. L. 1969. Flowering plants: origin and dispersal. Translated from Russian by C. Jeffrey. Oliver and Boyd, Edinburgh, Ireland.

TAKHTAJAN, A. L. 1980. Outline of the classification of flowering plants (Magnoliophyta). *Botanical Review* 46: 225-359.

TAKHTAJAN, A. L. 1997. Diversity and classification of flowering plants. Columbia University Press, New York, New York, USA.

THOMPSON, J. D., T. J. GIBSON, F. PLEWNIAK, F. JEANMOUGIN, AND D. G. HIGGINS. 1997. The Clustal X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research* 24: 4876-4882.

THORNE, R. F. 1968. Synopsis of a putatively phylogenetic classification of the flowering plants. *Aliso* 6: 57-66.

THORNE, R. F. 1983. Proposed new realignments in the angiosperms. *Nordic Journal of Botany* 3: 85-117.

THORNE, R. F. 1992. Classification and geography of the flowering plants. *Botanical Review* 58: 225-348.

THORNE, R. F. 2000. Classification and Geography of dicotyledons. *Botanical Review* 66: 441- 650.

TITMAN, P. W. 1949. Studies in the woody anatomy of the family Nyssaceae. *Journal of the Elisha Mitchell Scientific Society* 65: 245-261.

WAGENITZ, G. 1992. The Asteridae: evolution of a concept and its present status. *Annals of Missouri Botanical Garden* 79: 209-217.

WANGERIN, W. 1910. Das pflanzenreich. series IV, heft 41, A. Engler [ed.]. W. Engelmann, Weinheim/Bergstraße, Germany.

XIANG, Q.-Y. 1999. Systematic affinities of Grubbiaceae and Hydrostachyaceae within Cornales-insights from *rbcL* sequences. *Harvard Papers in Botany* 4: 527-542.

XIANG, Q.-Y., AND D. E. SOLTIS. 1998. *RbcL* sequence data defined a cornaceous clade and clarify relationships of Cornaceae *sensu lato*. *In* D. E. Boufford and H. Ohba [eds.], Sino-Japanese flora–its characteristics and diversification, 123-137. University of Tokyo, Tokyo, Japan.

XIANG, Q.-Y., M. MOODY, D. E. SOLTIS, C. FAN, AND P. S. SOLTIS. 2002. Relationships within Cornales and circumscription of Cornaceae: *matK* and *rbcL* sequence data and effects of outgroups and long branches. *Molecular Phylogenetics and Evolution* 24: 35-57.

XIANG, Q.-Y., D. E. SOLTIS, D. R. MORGAN, AND P. S. SOLTIS. 1993. Phylogenetic relationships of *Cornus* L. sensu lato and putative relatives inferred from *rbcL* sequence data. *Annals of Missouri Botanical Garden* 80: 723-734.

XIANG, Q.-Y., D. E. SOLTIS, AND P. S. SOLTIS. 1998. Phylogenetic relationships of Cornaceae and close relatives inferred from *matK* and *rbcL* sequences. *American Journal of Botany* 85: 285-297.

YANG, Z. 1996. Phylogenetic analysis using parsimony and likelihood methods. *Journal of Molecular Evolution* 42: 294-307.

ZANIS, M. J, D. E. SOLTIS, P. S. SOLTIS, S. MATHEWS, AND M. J. DONOGHUE. 2002. The root of the angiosperms revisited. *Proceedings of the National Academy of Sciences, USA* 99: 6848-6853.

ZHU, W.-H., AND Q.-B. XIANG. 1999. Morphological characters of the genus *Diplopanax* Hand.-Mazz. and its systematic implication. *Bulletin of Botanical Research* 19: 286-291.

Table 1. Comparison of taxonomic treatments of Cornales by different authors.

Harms (1898)	Wangerin (1910)	Hutchinson (1967)	Takhtajan (1980)	Cronquist (1981)	Dahlgren (1983)	Takahtajan (1997)	APG (1998)	Thorne (2000)
Umbelliflorae	Cornales	Araliales	Cornales	Cornales	Cornales	Cornales	Cornales	Cornales
Araliaceae	Alangiaceae	Alangiaceae	Alangiaceae	Alangiaceae	Adoxaceae, Alangiaceae	Alangiaceae	Cornaceae	Vitineae
Umbelliferae	Cornaceae	Araliaceae	*Alangium*	Cornaceae	Alseuosmiaceae	*Alangium*	*Alangium*	Vitaceae
Cornaceae	Curtisioideae	*Helwingia*	Aucubaceae	*Aralidium*	Anisophylleaceae	Cornaceae	*Camptotheca*	Vitoideae
Alangioideae	Cornoideae	Caprifoliaceae	Cornaceae	*Aucuba*	Aquifoliaceae	*Afrocrania*	*Cornus*	Leeoideae
Alangium	*Aucuba*	Cornaceae	Cornoideae	*Cornus*	Aralidiaceae	*Cornus*	*Curtisia*	Gunnerineae
Curtisioideae	*Cornus*	*Afrocrania*	Curtisioideae	*Corokia*	Aucubaceae	*Cynoxylon*	*Davidia*	Gunneraceae
Curtisia	*Corokia*	*Aucuba*	Mastixioideae	*Curtisia*	Caprifoliaceae	*Swida*	*Diplopanax*	*Gunnera*
Cornoideae	*Griselinia*	*Chamaepericlymenum*	Davidiaceae	*Griselinia*	Cardiopteridaceae	Curtisiaceae	*Mastixia*	Cornineae
Aucuba	*Helwingia*	*Cornus*	*Davidia*	*Helwingia*	Columelliaceae	*Curtisia*	*Nyssa*	Cornaceae
Cornus	*Kaliphora*	*Corokia*	Garryaceae	*Kaliphora*	Cornaceae, Davidiaceae	Davidiaceae	Grubbiaceae	*Cornus*
Corokia	*Melanophylla*	*Curtisia*	Griseliniaceae	*Mastixia*	Dulongiaceae	*Davidia*	Hydrangeaceae	Nyssaceae
Griselinia	*Toricellia*	*Cynoxylon*	Helwingiaceae	*Melanophylla*	Eremosynaceae	Mastixiaceae	Hydrostachyaceae	Davidioideae
Helwingia	Mastixioideae	*Dendrobenthamia*	Melanophyllaceae	*Toricellia*	Escalloniaceae	*Diplopanax*	Loasaceae	*Davidia*
Kaliphora	*Mastixia*	*Griselinia*	*Melanophylla*	Garryaceae	Garryaceae	*Mastixia*		Nyssoideae
Melanophylla	Garryaceae	*Kaliphora*	*Kaliphora*	Nyssaceae	Helwingiaceae	Nyssaceae		*Camptotheca*
Toricellia	Nyssaceae	*Mastixia*	Nyssaceae	*Camptotheca*	Hydrangeaceae	*Camptotheca*		*Nyssa*
Davidioideae	Davidioideae	*Melanophylla*	*Camptotheca*	*Davidia*	Icacinaceae	*Nyssa*		Mastixioideae
Davidia	*Davidia*	*Swida*	*Nyssa*	*Nyssa*	Montiniaceae, Nyssaceae			*Diplopanax*
Garryoideae	Nyssoideae	*Toricellia*	Toricelliaceae		Paracryphiaceae			*Mastixia*
Mastixioideae	*Camptotheca*	Garryaceae			Phellinaceae			Curtisiaceae
Nyssoideae	*Nyssa*	Nyssaceae			Pterostemonaceae			*Curtisia*
Camptotheca		*Camptotheca*			Sambucaceae			Alangiaceae
Nyssa		*Davidia*			Sphenostemonaceae			*Alangium*
		Nyssa			Stylidiaceae			
					Symplocaceae			
					Tetracarpaeaceae			
					Torricelliaceae			
					Tribelaceae, Viburnaceae			

Table 2. Sources of species sampled in the study of 26S rDNA sequencing of Cornales.

Species	Sources and location of vouchers	GenBank accession no.
Alangium		
A. platanifolium (Sieb. & Zucc.) Harms	Soltis 2543, Japan.	AF297544
A. kurzii Craib	Xiang 02-72, China, 2002.	AY260007
A. chinense (Lour.) Harms.	Deng16101, Xinning, Hunan, China, 2001. South China Institute of Botany.	AY260008
A. chinense (Lour.) Harms.	Deng 15866, Boguo, Guangdong, China. South China Institute of Botany.	AY260009
Camptotheca acuminata Decne.	Strybing Arboretum #74-180	AY260010
Cornus		
C. canadensis L.	Xiang et al. 198, WS.	AF297530
C. controversa Hemsl.	Arnold Arboretum 20458, WS.	AF297541
C. disciflora Sesse & Moc. Ex DC.	R. Fernandez N., 5041, Mexico	AY260011
C. florida L.	Xiang 250, WS.	AF297532
C. kousa Hance	Xiang 310, Ohio State University campus.	AF297533
C. mas L.	Arnold Arboretum 577-51-A, WS.	AF297535

C. oblonga Wall.	Sun, s.n., Bot. Gard. Kunming, China	AF297539
C. officinalis Seib. & Zucc.	Boufford et al. 26065, GH.	AF297536
C. racemosa Lam.	Xiang et al. 157, WS.	AF297538
C. sessilis Torr. Ex Durand	Terry M. Hardig, California 1994.	AF297537
C. suecica L.	Chris Brochmann, 94-388, Norway.	AF297531
C. unalaschkensis Ledeb.	Xiang 210, WS.	AF297534
C. volkensii Harms	Knox 2528, Africa.	AF297542
C. walteri Wangerin	Arnold Arboretum 414-67-1, WS.	AF297540
Curtisia dentata (Burm.) G. A. Sm.	Edwards 918, NU	AY260012
Davidia involucrata Baill.	U.S. National Arboretum, #12067	AY260013
Diplopanax stachyanthus Hand. – Mazz.	J. Li, 2001, Ruyang, Guangdong, China	AY260014
Grubbia		
G. rosmarinifolia Berg.	M. W. Chase 5706K	AY260019
G. tomentosa (Thunb.) Harms	M. W. Chase 5805K	AY260020
Hydrostachys		
H. angustisecta Engl.	Bremer 3089 Steven & Mariette Mariette Manktelow	AY260022
H. angustisecta Engl.	HS-4, Henrik Strandgaard, 1992, Malawi	AY260026
H. imbricata A.Juss.	Madagascar, Schatz et al. 3414	AY260023

H. insignis Mildbr. & Reim	HS-3, Henrik Strandgaard, 1992, Malawi	AY260027
H. multifida A. Juss.	Madagascar, Schatz et al. 3413	AY260021
H. polymorpha Klotzsch	HS-2, Henrik Strandgaard, 1992, Malawi	AY260024
Hydrostachys sp.	Peer Hansen, 1993, Madagascar	AY260025
Mastixia		
M. caudatilimba C. Y. Wu ex Soong	Zan-He Ji, s.n., WS	AY260015
M. eugenioides K.M.Matthew	Zhu & Wang 3002, 1991, Yunnan, China	AY260017
M. pentandra subsp. Chinensis (Merrill) K, M, Matthew	D2276- 1630, 1997, China	AY260016
Nyssa		
N. ogeche Marsh.	U.S. National Arboretum, s.n.	AF297545
N. sylvatica Marsh.	C. Fan 02-39, 2002	AY260018
Hydrangeaceae		
Broussaisia arguta Gaud.	Hawaii, Flynn 5060	AY260039
Cardiandra alternifolia Sieb. Et Zucc.	Honshu, Japan; Fujii 1992, (OSA)	AY260040
Decumaria sp.	Chase et al. (1993)	AY260043
Deutzia rubens Rehder	Arnold Arboretum #1003-86-Mass.	AY260034

Fendlera rupicola Engelm. et Gray	Soltis et al., 1995	AY260041
Fendlerella utahensis (Wats.) Heller	Okane 3023 (UC)	AY260035
Hydrangea arborescens L.	C. Fan 02-29, 2002	AY260032
Jamesia americana T. & G.	Soltis et al., 1995	AY260042
Philadelphus caucasicus Koehne	Xiang 307	AY260037
Philadelphus hirsutus Nutt.	Arnold Arboretum #320-79-4	AY260036
Pileostegia sp.	Taiwan, Qiu, W. L. 1992	AY260038
Platycrater arguta Sieb. & Zucc.	2 June, 1992 Makota op	AY260033
Schizophragma hydrangeoides Sieb. et Zucc.	Soltis 2516, Japan.	AF297543
Loasaceae		
Eucnide urens (A. Gray) Parry	Hufford 552, WS	AY260031
Mentzelia decapetala (Pursh) Urb. & Gilg	Linda Cook 499, 7 Aug. 1996	AY260030
Petalonyx nitidus S. Wats	Hufford 554, DUL	AY260028
Petalonyx parryi A.Gray	Hufford 2011, 26 May, 1997	AY260029
Outgroups		
Apium graveolens L.	Genbank	AF479195
Corokia cotoneaster Raoul	Genbank	Af479187
Fouquieria columnaris Kellogg	Genbank	Af479159
Garrya elliptica Dougl.ex Lindl.	Genbank	AF479181
Halesia diptera L.	Genbank	AF479157
Ilex opaca	Genbank	AF479203

Myoporum mauritianum A.DC.	Genbank	AF479170
Panax quinquefolius L.	Genbank	AF479193
Petroselinum crispum (Mill.) A.W.Hill	Genbank	AF479237
Sarracenia purpurea L.	Xiang 252, WS	AY260044
Shortia galacifolia Torr. & Gray.	C. Fan 02-44, 2002	AY260045
Solanum lycopersicum L.	GenBank	X13557
Styrax japonicus Sieb. & Zucc.	Genbank	AF479156
Tragopogon dubius Scop.	Genbank	AF036493
Veronica anagallis-aquatica L.	Genbank	AF479169

Table 3. ILD test between 26S rDNA and *matK-rbcL* data sets with different major lineages excluded and included alone.

Lineages excluded/included alone	Sum of tree lengths for original partition (excluded/included alone)	p value (excluded/included alone)
Cornus-Alangium	4018/589	0.001**/0.121
Nyssoids-mastixioids	4215/401	0.001**/0.08
Loasaceae	4335/N.A.	0.001**/N.A.
Hydrangeaceae	3804/815	0.003**/0.001**
Hydrangeeae	4088/359	0.001/0.001**
Curtisia-Grubbia	4411/N.A.	0.001**/N.A.
Curtisia-Grubbia-Loasaceae	4144/542	0.001**/1.0
Hydrostachyaceae	4223/N.A.	0.001**/N.A.
Hydrostachyaceae + outgroups	2503/2054	0.001**/0.01**
Outgroups	2965/1594	0.001**/0.22

Note: **Significant discordances between partitions of 26S rDNA and *matK-rbcL*

N.A.: not applicable due to ModelTest required at least four taxa.

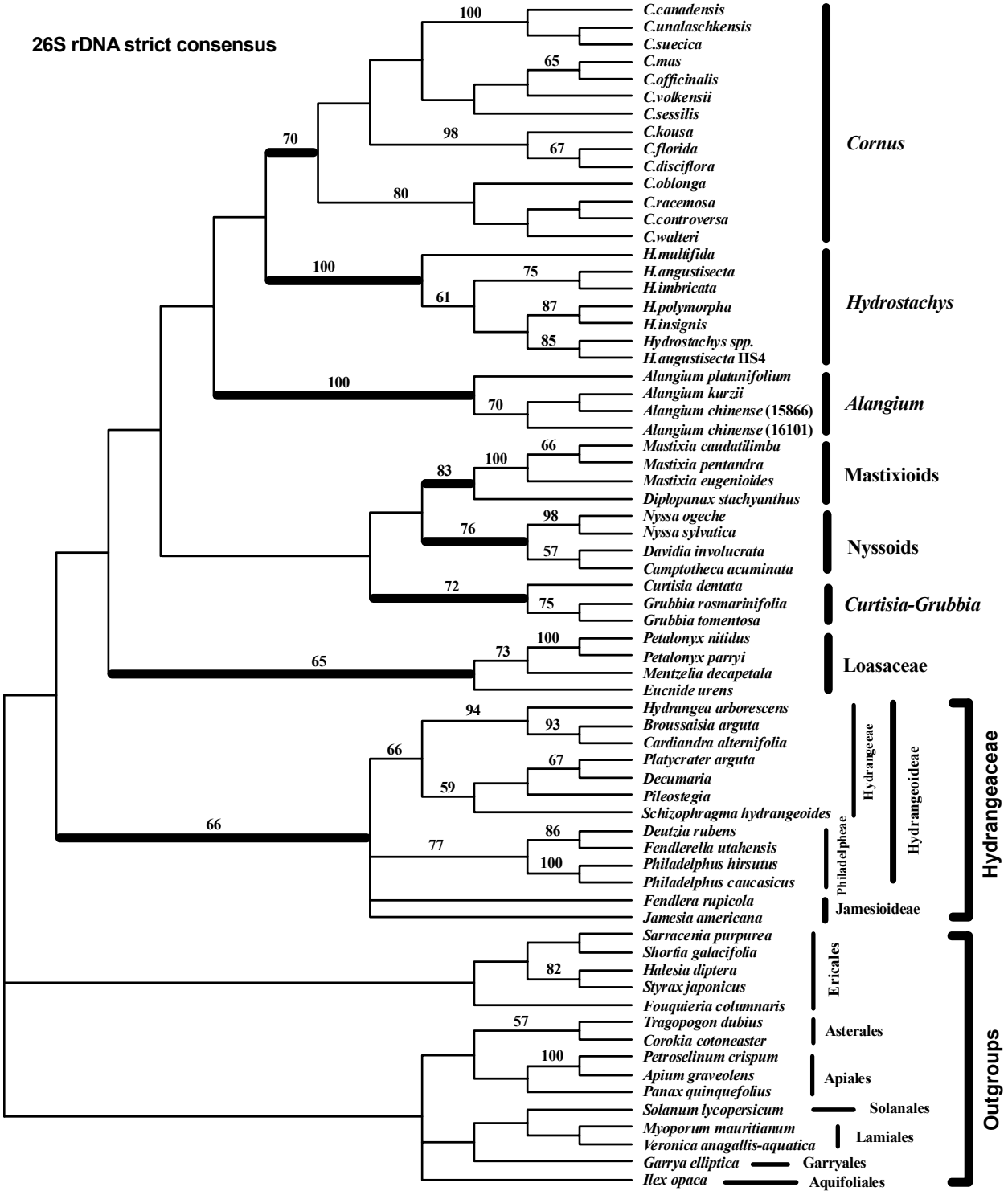


Fig. 1. The strict consensus tree from parsimony analysis of 26S rDNA sequences. Bootstrap values (>50%) are indicated above branches. The major clades are marked by thickened lines.

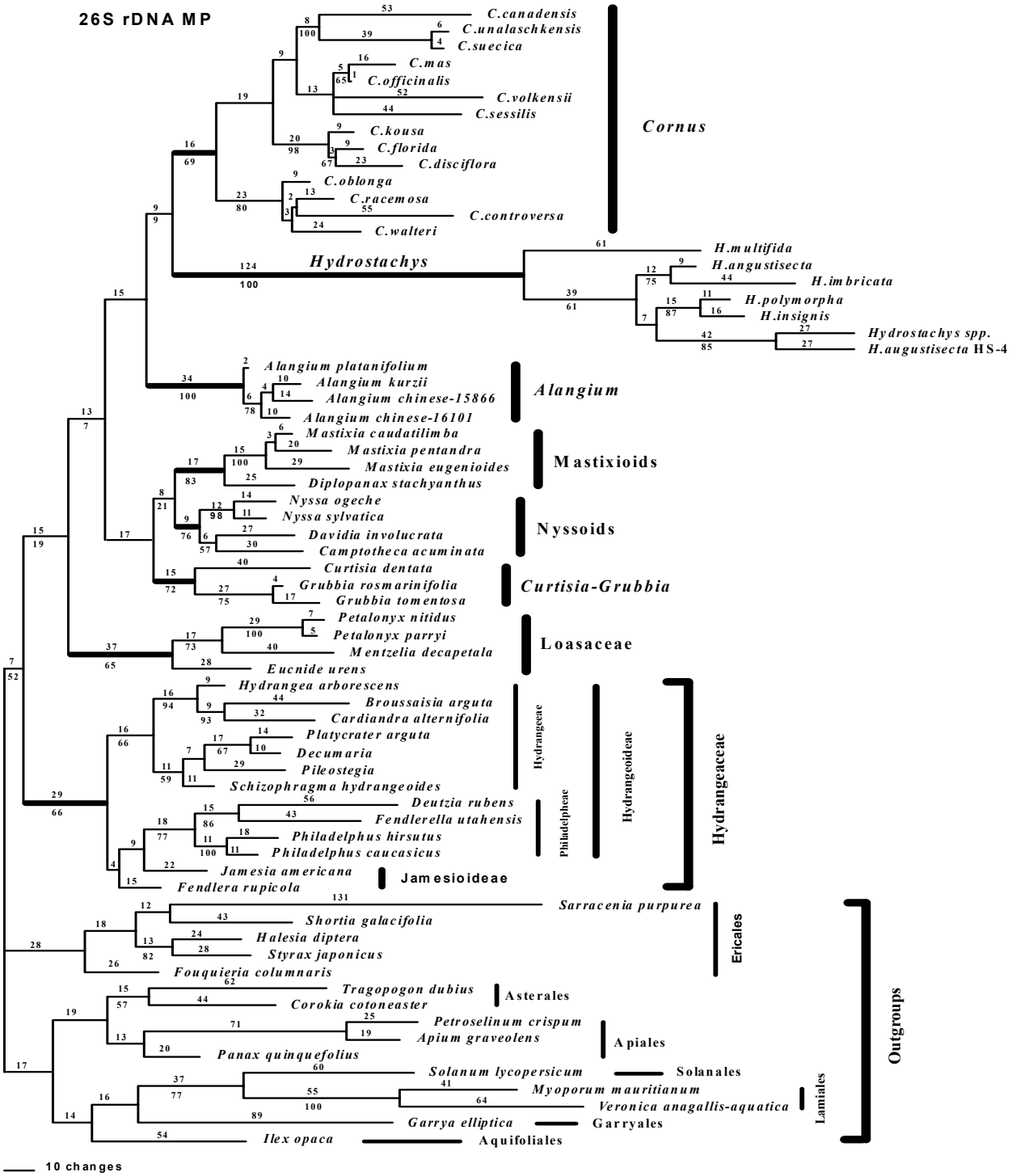


Fig. 2. One of 47 equally parsimonious trees from parsimony analysis of 26S rDNA sequences (tree length = 2947 steps, CI = 0.487 excluding uninformative characters, RI = 0.632). Base substitutions are indicated above branches; bootstrap values (>5%) are indicated below branches. The major clades are marked by thickened lines.

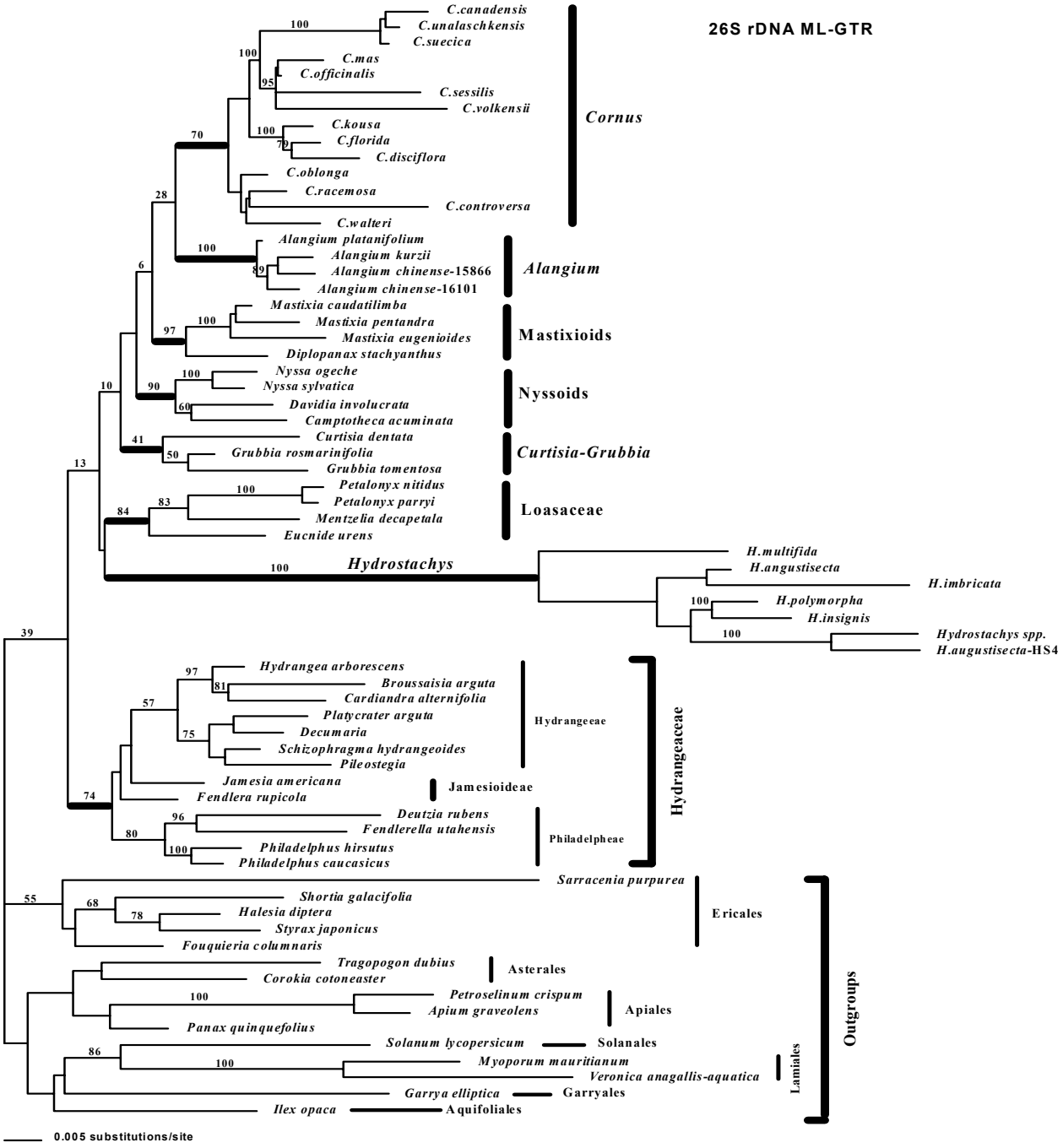


Fig. 3. The maximum likelihood tree from analysis of the 26S rDNA sequences using GTR + I + Γ model with the following parameter values: rate matrix of R with AC = 1.094, AG = 2.342, AT = 1.599, CG = 1.103, CT = 7.888; base frequencies = A: 0.235, C: 0.247, G: 0.314, T: 0.203; proportion of invariable sites = 0.488; α of gamma distribution = 0.522. Bootstrap values (>5%) obtained using neighbor joining bootstrap analysis employing ML distance (detailed see Materials and Methods) are indicated above branches (-Ln likelihood = 21451.620). The major clades are marked by thickened lines.

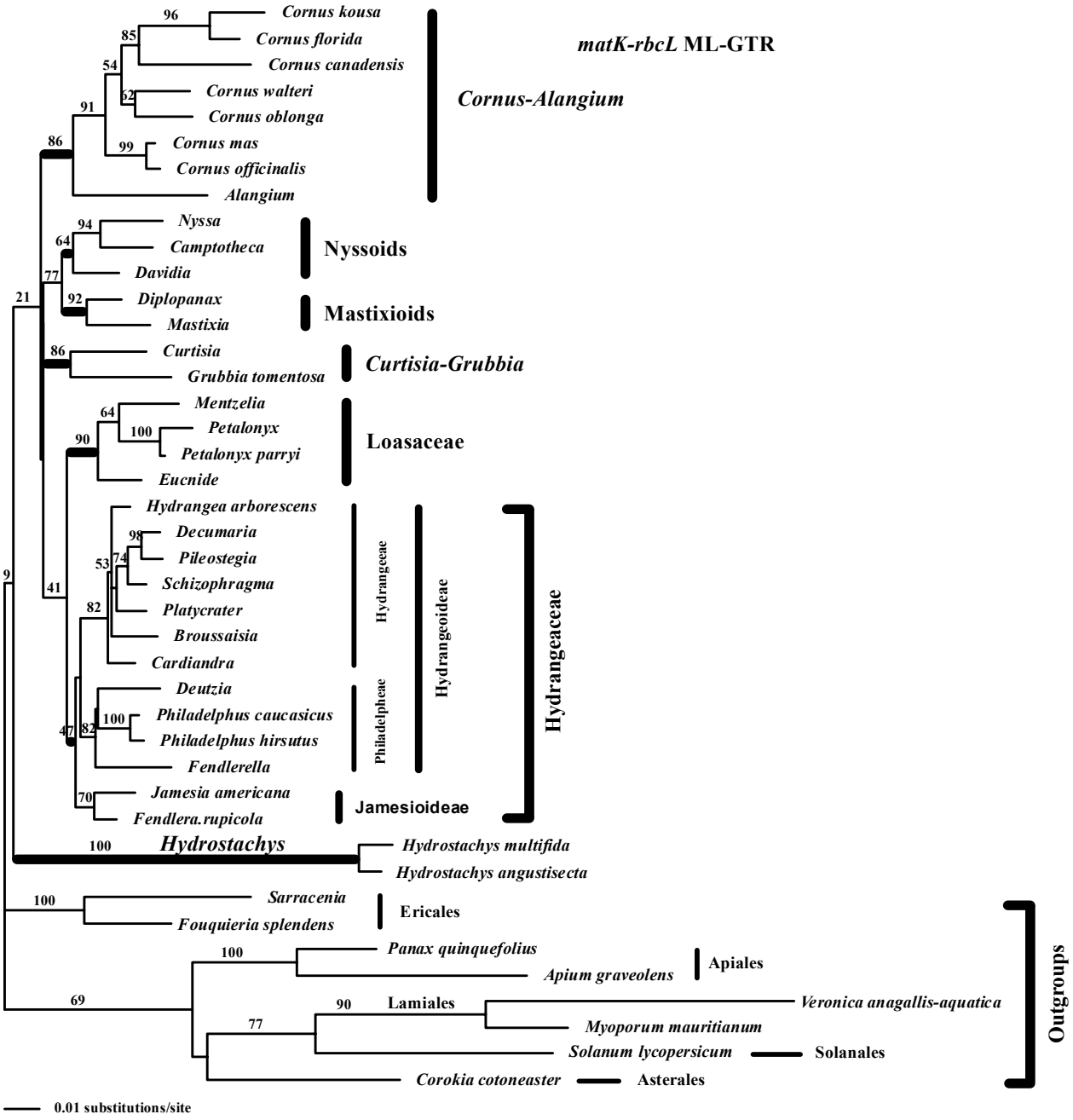


Fig. 4. The maximum likelihood tree from analysis of combined *matK* and *rbcL* sequence of 42 taxa also having 26S rDNA sequences using GTR + I + Γ model with parameter values: rate matrix of R with AC = 1.389, AG = 2.274, AT = 0.393, CG = 0.919, CT = 2.274; base frequencies = A: 0.296, C: 0.187, G: 0.189, T: 0.328; proportion of invariable sites = 0.384; α of gamma distribution = 1.256. ML Bootstrap values (>5%) are indicated above branches (-Ln likelihood = 17850.031). The major clades are marked by thickened lines.

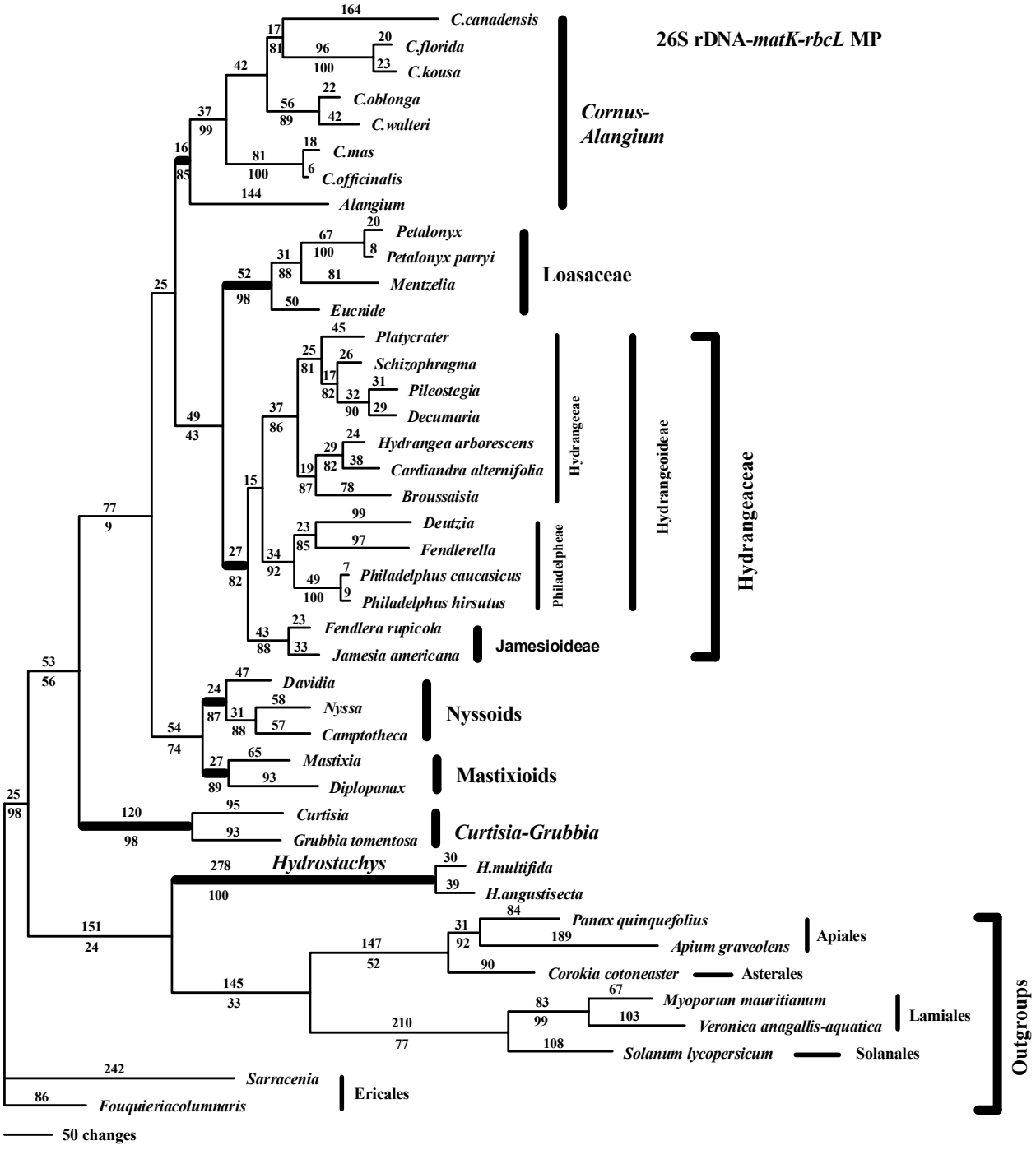


Fig. 5. The single most parsimonious tree from analysis of combined 26S rDNA-*matK-rbcL* sequence data (tree length = 4735 steps, CI = 0.570 excluding uninformative characters, RI = 0.547). Base substitutions are indicated above branches; bootstrap values (>5%) are indicated below branches. The major clades are marked by thickened lines.

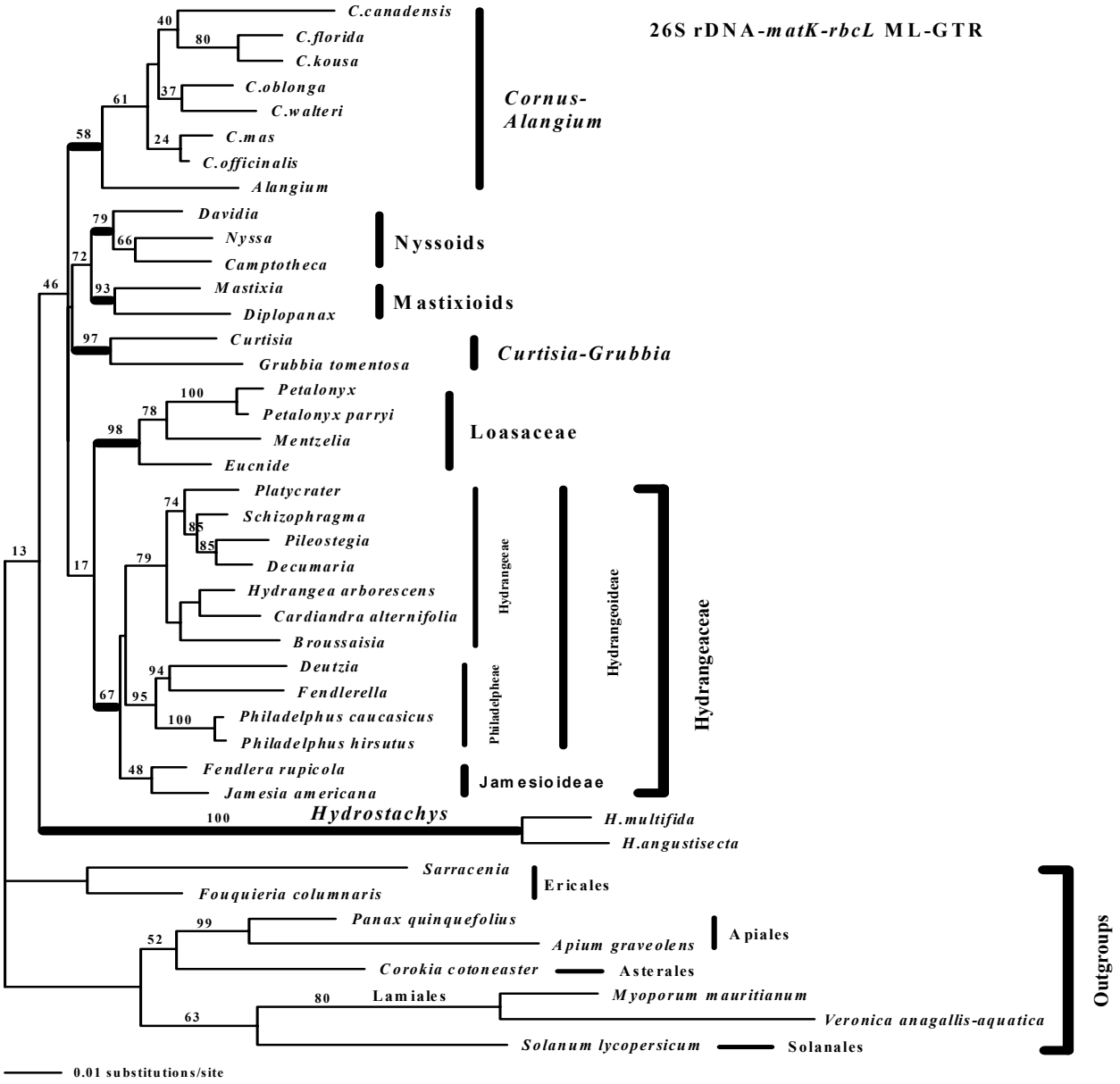


Fig. 6. The maximum likelihood tree from analysis of combined 26S rDNA-*matK-rbcL* sequence data using the GTR + I + Γ model with parameter values: rate matrix of R with AC = 1.279, AG = 2.048, AT = 0.739, CG = 0.958, CT = 4.287; base frequencies = A: 0.265, C: 0.219, G: 0.257, T: 0.258; proportion of invariable sites = 0.498; α of gamma distribution = 0.735. Bootstrap values (>5%) approximated using neighbor joining employed with ML distance (see Materials and Methods) are indicated above branches (-Ln likelihood = 34716.487). The major clades are marked by thickened lines.

Chapter II

PHYLOGENETIC RELATIONSHIPS WITHIN *CORNUS* (CORNACEAE) BASED ON 26S rDNA SEQUENCES[1]

CHUANZHU FAN[2] AND (JENNY) QIU-YUN XIANG

Department of Botany, North Carolina State University, Campus Box 7612, Raleigh, North Carolina 27695-7612, USA

Published in *American Journal of Botany* 2001, 88(6): 1131-1138.

Chuanzhu did all of the work reported in this paper, but Dr. Jenny Xiang provided scientific advice and guidance.

[1]Manuscript received_____; revision accepted_____.

The study was supported by the Research Office of Idaho State University. Our special thanks go to Dr. Larry Smith and Erin O'Leary-Jepsen at the Molecular Research Core Facility at ISU for running the sequencing gels, to Dr. Zack Murrell for providing DNAs of *C. volkensii*, to Dr. Chris Brochmann for collecting plant material of *C. suesica*, to Dr. D. E. Boufford for collecting *C. officinalis*, to Dr. Gregory M. Plunkett and Dr. Robert K. Kuzoff for their insightful review comments on the manuscript.

[2]Author for reprint requests (Tel: 919-515-3345; fax: 919-515-3436; email: cfan3@unity.ncsu.edu).

Abstract: Phylogenetic relationships within the dogwood genus *Cornus* have been highly controversial due to the great morphological heterogeneity. Earlier phylogenetic analyses of *Cornus* using chloroplast DNA (cpDNA) data (including *rbcL* and *matK* sequences, as well as restriction sites) and morphological characters suggested incongruent relationships within the genus. The present study generated sequence data from the nuclear gene 26S rDNA for *Cornus* to test the phylogenetic hypotheses based on cpDNA and morphological data. The 26S rDNA sequence data obtained represents 16 species, 13 from *Cornus* and three from outgroups, having an aligned length of 3380 bp. Both parsimony and maximum likelihood analyses of these sequences were conducted. Trees resulting from these analyses suggest relationships among subgroups of *Cornus* consistent with those inferred from cpDNA data. That is, the dwarf dogwood (subg. *Arctocrania*) and the big-bracted dogwood (subg. *Cynoxylon* and subg. *Syncarpea*) clades are sisters, which are, in turn, sister to the cornelian cherries (subg. *Cornus* and subg. *Afrocrania*). This red-fruited clade is sister to the blue- or white-fruited dogwoods (subg. *Mesomora*, subg. *Kraniopsis,* and subg. *Yinquania*). Within the blue- or white-fruited clade, *C. oblonga* (subg. *Yinquania*) is sister to the remainder, and subg. *Mesomora* is sister to subg. *Kraniopsis*. These relationships were also suggested by the combined 26S rDNA and cpDNA data, but with higher bootstrap and Bremer support in the combined analysis. The 26S rDNA sequence data of *Cornus* consist of 12 expansion segments spanning 1034 bp. These expansion segments evolve approximately four times as fast as the conserved core regions. The study provides an example of phylogenetic utility of 26S rDNA sequences below the genus level.

Key words: 26S rDNA; combining data; Cornaceae; *Cornus* ; cpDNA; molecular evolution; molecular phylogeny.

INTRODUCTION

Cornus L. sensu lato consists of ~ 55 species that are mostly trees and shrubs, and rarely perennial herbs with woody rhizomes. The genus is widely distributed in the northern hemisphere, with centers of diversity in eastern Asia, eastern North America, and western North America. Two species of the genus are endemic to South America and one species is endemic to tropical Africa (see Table 1). Species of *Cornus* are morphologically diverse. For example, various inflorescences are found within the genus, including open compound cymes with minute, non-modified bracts; umbellate cymes with four basal, scale-like bracts; capitate cymes subtended by four large showy bracts; and minute compound cymes with four showy bracts. The color of fruits also varies among species (see Table 1). Due to the great morphological diversity encompassed by the Linnaean circumscription of *Cornus*, the taxonomy and phylogenetic relationships of subgroups within the genus are highly controversial (see Eyde, 1988; Murrell, 1993; Xiang et al., 1996). The genus has been divided into several distinct genera (Hutchinson, 1942; Nakai, 1909; Rafinesque, 1838; Spach, 1839; and von Berchtold and Opiz, 1838; Pojarkova, 1950; also see Murrell, 1993), or into various numbers of subgenera (Ferguson, 1966; Harms, 1898; Wangerin, 1910; Xiang, 1987; Eyde, 1987, 1988). Following the broad view of *Cornus*, a total of ten subgenera have been recognized by different authors at one time or another (see Table 1; Ferguson, 1966; Xiang, 1987; Eyde, 1988; Murrell, 1993).

In order to understand relationships within *Cornus*, several sets of data have been recently collected for the genus for phylogenetic analyses. These include molecular data from the chloroplast genome (Xiang et al., 1996; Xiang, Soltis, and Soltis, 1998) and

morphological characters (Murrell, 1993). Phylogenetic analyses of these data identified five major lineages within *Cornus*: (1) *C. oblonga*, an enigmatic blue-fruited dogwood, (2) other blue- or white-fruited dogwoods, (3) the cornelian cherries, (4) the big-bracted dogwoods, and (5) the dwarf dogwoods (the herbaceous species). However, relationships among these groups suggested by cpDNA data are different from those suggested by morphological data. The cpDNA data (combined *rbcL* and *matK* sequences and restriction sites) suggested that the dwarf dogwoods are sisters of the big-bracted dogwoods with the cornelian cherries sister to them, and *C. oblonga* is the sister of the remainder of blue- or white-fruited dogwoods (see Fig. 1; Xiang et al., 1996; Xiang, Soltis, and Soltis, 1998). These results are consistent with the hypothesis of Eyde (1988) based on synthesis of information available without performing phylogenetic analyses. In contrast, cladistic analysis of 28 morphological, anatomical, chemical, and cytological characters of *Cornus* by Murrell (1993) suggested that the cornelian cherries and the big-bracted dogwoods are sister groups, and the dwarf dogwoods are sister to them, with *C. oblonga* being the first-branching lineage in *Cornus* (see Fig. 2; Murrell, 1993; Hardin and Murrell, 1997). To explore possible sources of incongruence between the cpDNA and morphological data, additional phylogenetic analyses, especially analyses of molecular data from the nuclear genome, are necessary.

Although the most widely used nuclear phylogenetic markers below the generic level have been the ITS (internal transcribed spacer) sequences of ribosomal genes (see reviews by Baldwin et al., 1995; Soltis and Soltis, 1998), our initial analyses within *Cornus* revealed a high level of ITS sequence divergence among species from the five major lineages, which made unambiguous alignment of sequences unfeasible (Xiang and Fan, unpublished data). Therefore, we turned to a more slowly evolving nuclear gene 26S rDNA. Recent studies

have demonstrated the great potential of 26S rDNA sequences in elucidating phylogenetic relationship at various taxonomic levels of seed plants (at and above the generic level) (e.g. Mishler et al., 1994, used rRNA sequences only; Ro, Keener, and McPheron, 1997; Kuzoff et al., 1998; Stefanovic et al., 1998; Soltis and Soltis, 1998; Ashworth, 2000). In our on-going study of 26S rDNA sequencing for Cornales, several species representing the five different subgroups of *Cornus* were included. Our preliminary results revealed that sufficient sequence variation exists among *Cornus* species. Thus, we employed comparative 26S rDNA sequencing to reconstruct a nuclear phylogeny of *Cornus*. The goals of this study were: (1) to determine phylogenetic relationships among subgroups of *Cornus* using nuclear 26S rDNA sequences; (2) to test the phylogenetic hypotheses based on morphology and cpDNA data; (3) to gain insights into the evolution of some subgroup-dignostic morphological characters in *Cornus*.

MATERIALS AND METHODS

Sampling - - Thirteen species of *Cornus* L. representing the range of morphological diversity of the genus and the five major lineages identified through earlier phylogenetic analyses of cpDNA were sampled in the 26S rDNA sequencing study (Table 2; also see Table 1). All subgenera except subg. *Discocrania* and subg. *Sinocornus* were included. These two subgenera, each represented by a single species, were not sampled because of lack of DNA. Previous cpDNA studies indicated that *C. disciflora* (the only member of subg. *Discocrania*) is a member of the big-bracted dogwoods and *C. chinensis* (the only member of subg. *Sinocornus*) is a member of the cornelian cherries (see Table1, Fig. 1; Xiang et al.,

1996; Xiang, Soltis, and Soltis, 1998). Thus, even without these two species our sampling well represented all major subgroups within the genus. Although subgenus *Kraniopsis* is the largest subgenus in *Cornus* (~30 spp), all species in the subgenus are morphologically very similar, and they formed a strongly supported monophyletic group in the cpDNA study (Xiang et al., 1996). Therefore, only two species were sampled from this subgenus. All of the species sampled were included in previous *matK* and *rbcL* sequencing and chloroplast DNA restriction site analyses, except *C. suecica* (a dwarf dogwood) and *C. volkensii* (the single species from Africa). These two species had not been included in the cpDNA studies because of the lack of materials of these taxa at the time the studies were completed. Three genera, *Nyssa*, *Alangium,* and *Schizophragma* (Hydrangeaceae), were chosen as the outgroups for the present study based on the results of broad phylogenetic analyses of *matK* amd *rbcL* sequences for Cornaceae (Xiang, Soltis, and Soltis, 1998), which suggested that *Alangium* is the sister of *Cornus*; *Nyssa* and Hydrangeaceae are close relatives of *Cornus*.

DNA isolation, PCR amplification and DNA sequencing of 26S rDNA-- To maximize comparability between the present and previous studies, DNAs used herein were those isolated for previous *rbcL* and *matK* sequencing studies. The procedures of isolating DNA were described elsewhere (Xiang et al., 1993; Xiang, Soltis, and Soltis, 1998). The 26S rDNA sequences were amplified via PCR (polymerase chain reaction) from total DNA aliquots using the forward primer N-nc26S1 (5'-CGACCCCAGGTCAGGCG-3') and the reverse primer 3331rev (5'-ATCTCAGTGGATCG TGGCAG-3') following Kuzoff et al. (1998) with slight modifications. Our PCR reactions contained the following: 5 μL of 10x Mg free buffer, 6 μL of 25mmol/L $MgCl_2$, 10 μL of 2.5μmol/L dNTP, 1.0 μL of 20μmol/L N-nc26S1 (forward primer), 1.0μL of 20μmol/L 3331 rev (reverse primer), 5μL of

DMSO(dimethyl sulfoxide), 0.3μL of *Taq* polymerase (Promega), 2.0μL of 20ng/μL total DNA extract, and 19.7μL of deionized water. The PCR reaction mix was covered with two drops of mineral oil, and run on a Robocycle PCR machine as the following: (1) 94°C for 3 min for one cycle; (2) 30 cycles of 94°C for 1 min, 55°C for 1 min, 72°C for 3.5 min; (3) a terminal phase at 72°C for 5 min.

The double-stranded (DS) PCR products were subsequently purified via precipitation with 60 μL of 20% PEG 8000/2.5mol/L NaCl on ice for at least 1 h (Soltis and Soltis, 1997). The precipitated DS products were centrifuged for 15 min at 14000 rpm at 4°C. The DNA pellets were then washed with 1000 μL of 75% ethanol (prechilled to 4°C) and centrifuged for 3 min. The DNA pellets were washed again with chilled 95% ethanol and centrifuged for 3 min at 4°C. The ethanol was decanted, and the pellets were dried in a vacuum. The dried DNA products were resuspended in 20-30 mL of ddH_2O. One microlitre of the clean PCR products was electrophoresed in a 1% agarose mini-gel for quantification.

The purified ds DNA products were used as the templates for sequencing on an ABI-377 automated sequencer following the standard protocol recommended by the company (Applied Biosystems, Foster City, CA94404, USA). For some species, DMSO was added to the sequencing reactions to obtain clear sequences. Sixteen sequencing primers (N-nc26S1, N-nc26S3, N-nc26S4, N-nc26S5, N-nc26S6, N-nc26S8, N-nc26S10, N-nc26S12, N-nc26S14, 268rev, 641rev, 950rev, 1449rev, 2134rev, 2782rev, and 3331rev) described in Kuzoff et al. (1998) were used to obtain the entire sequence of 26S rDNA. The sequence chromatogram output files for all species were checked and edited base by base before being aligned manually.

Phylogenetic analysis **- -** The 26S rDNA sequences representing 16 taxa comprise 3380 aligned bp with small gaps (see Results). The data matrix was analyzed with both parsimony and maximum likelihood (ML) methods using PAUP 4.0b4 (Swofford, 2000). For parsimony analysis, gaps were coded as missing data; branch-and-bound search was conducted. Branch-and-bound search was performed using furthest taxon addition sequence and initial upper bound computed via stepwise addition. To evaluate relative robustness of the clades found in the most parsimonious trees, a bootstrap and decay analyses were conducted. Bootstrap analysis (Felsenstein, 1985) of 1000 replicates was conducted using fast heuristic search and TBR branch-swapping. The method for decay analysis (Bremer, 1988) was described in Xiang et al. (1996), which followed Eernisse and Kluge (1993). This method involves examining each node of interest in turn using a constraint statement that specifies only the node of interest being monophyletic and saving the shortest trees that do not satisfy this criterion. The difference between the length of these trees and the true shortest trees is used as the decay value for that node.

ML analyses were first conducted using heuristic searches with random taxon addition of five replicates with default settings of the ML program (i.e., empirical base frequency, HKY model requested [Hasegawa-Kishino-Yano, 1985] two-parameter model variant for unequal base frequency, ti/tv = 2, and equal rates for all sites). The resulting tree was simultaneously used to estimate values for base frequency, ti/tv ratio, proportion of invariable sites, and shape of the discrete gamma distribution of rates across sites, through maximum likelihood. A subsequent ML analysis using the estimated values of these parameters (A = 0.220264, C = 0.272649, G = 0.307582, T = 0.199505; ti/tv ratio = 1.343463 [kappa = 2.741222]; proportion of invariable sites = 0.704318; shape value for discrete gamma

distribution = 0.848768) was then conducted to see whether adjusting these parameters to the estimated values resulted in significant differences in tree topologies.

Since the results of analyses of 26S rDNA sequences suggested phylogenetic relationships within *Cornus* are highly congruent with those achieved via cpDNA data (combined *matK*, *rbcL* sequences and restriction sites), the 26S rDNA sequences were combined with cpDNA data for further parsimony and ML analyses to obtain a comprehensive view of relationships. The combined data matrix included ten species of *Cornus* and two outgroups (*Alangium* and *Nyssa*), each of 6338 characters, among which 3380 bp were from 26S rDNA, 1212 bp from *matK,* 1504 bp from *rbcL*, and 242 from restriction sites. Phylogenetic analyses of the combined 26S rDNA and cpDNA were conducted as above.

RESULTS

Sequence divergence **- -** The 26S rDNA sequences generated for the 16 species of *Cornus* varied from 3340 to 3370 bp in length before alignment. All sequences can be aligned easily by sight against the reference sequences from two angiosperm species, *Saxifraga mertensiana* Bong. (AFO36498, Kuzoff et al., 1998), and *Tragopogon dubius* Scop. (AFO36493, Kuzoff et al., 1998), obtained from GenBank. The aligned 26S rDNA sequences in *Cornus* and outgroups contained a total length of 3380 bp. Sequences for all species are complete except for *C. cotroversa* and *C. sessilis*. In *C. cotroversa,* 166 bp (bases 1542-1707) are missing and in *C. sessilis,* 476 bp (bases 1217-1692) are missing. For *C. controversa*, this region can not be amplified using PCR; for *C. sessilis*, although this region

can be amplified, we can not get a clean sequence from some primers. Despite repeated efforts, these missing data in these two species could not be obtained due to potential primer divergence. Among the 16 sequences of *Cornus* and outgroups, 391 of the 3380 sites are variable (11.56 %) and 137 sites (4.05%) are phylogenetically informative.

The alignment sequences of *Cornus* and outgroups required the addition of 20 small alignment gaps (1-5 bp in length). Eleven of these gaps appear to be autapomorphies, and nine of them are potentially phylogenetically informative within *Cornus* (Table 3, also see Fig. 3).

It is noteworthy that a majority of these indels (15 of 20) occur in the expansion segments, regions of gene that evolve more rapidly (Clark et al., 1984; Dover and Flavell, 1984; Flavell, 1986). The location of expansion segments of 26S rDNA sequences appears to be highly conserved over a wide range of taxa (Bult, Sweere, and Zimmer, 1995). According to the coordinates for expansion segment positions in the sequence of *Oryza sativa* (Kuzoff, et al. 1998), 12 expansion segments were identified in the 26S rDNA sequence data of *Cornus* and outgroups (Table 4). These expansion segments of 26S rDNA in *Cornus* and outgroups span a total of 1034 bp, among which 233 bp (22.53%) are variable and 97 sites (9.38%) are phylogenetically informative. These values are much higher than those for the conserved core regions, which contain 2346 bp, of which 158 (6.73%) are variable and have only 40 sites (1.71%) that are phylogenetically informative.

The average G+C content of the entire 26S rDNA (58.1%) and the conserved core regions (52.6%) is high. The expansion segments also have a higher average G+C content (68.5%).

Phylogenetic relationships inferred from 26S rDNA sequences - - Phylogenetic analyses of the entire 26S rDNA sequences of *Cornus* using parsimony found a single shortest tree of 577 steps (CI [consistency index] = 0.780, excluding the uninformative characters, RI [retention index] = 0.619; Fig. 3); this single shortest tree is completely resolved. Major clades (*C. oblonga* was a clade distinct from other blue-fruited *Cornus* in the earlier studies, and this why four clades, not five, are recovered here) identical to those recognized by earlier phylogenetic analyses of cpDNA data were identified in the 26S rDNA tree: (1) the blue- or white-fruited dogwoods (represented by *C. racemosa*, *C. walteri*, *C. controversa,* and *C. oblonga*); (2) the cornelian cherries (*C. mas*, *C. officinali*s, *C. sessilis* , and *C. volkensii*); (3) the big-bracted dogwoods (represented by *C. florida* and *C. kousa*); and (4) the dwarf dogwoods (*C. canadenisis*, *C. suecica*, and *C. unalaschkensis*). The dwarf dogwoods and the big-bracted dogwoods (all producing showy bracts on inflorescence) were recognized as sisters. This showy-bracted dogwood clade was, in turn, sister to the cornelian cherries. All of the blue- or white-fruited dogwoods formed a monophyletic group sister to the large clade consisting of the dwarf dogwoods, big-bracted dogwoods, and the cornelian cherries. Although all of the major clades (except the cornelian cherries) are strongly supported (with > 82% bootstrap value and > 6 decay index), the relationships among major groups described above are not strongly supported (with bootstrap value < 50 and decay index < 3) (Fig. 3). Within the dwarf dogwood lineage (with all of the three species sampled), *C. canadensis* and *C. unalaschkensis* are sisters. Within the cornelian cherry group (four of the five species were sampled, see Table 1), *C. mas* and *C. officinalis* are sisters, with *C. sessilis* sister to them, and *C. volkensii* is at base within the clade. Within the blue- or white-fruited lineage, *C. oblonga* is sister to the remainder of the lineage (Fig. 3).

The ML analysis using default setting (see Materials and Methods) found a tree with the best score of -Ln = 8531 with a topology identical to the parsimony tree (Fig. 4). The ML analysis using estimated values found a tree with a higher likelihood (-Ln = 8219), but showing the same relationships as described above.

Analyses of combined cpDNA and 26S rDNA sequence data -- The combined 26S rDNA and cpDNA data set contains 939 (14.82%) variable sites of which 347 are from 26S rDNA, 232 from *matK*, 153 from *rbcL*, and 207 from restriction sites; and 327 (5.16%) are phylogenetically informative sites, of which 98 are from 26S rDNA, 59 from *matK*, 65 from *rbcL*, and 105 from restriction sites. Parsimony analysis of the combined data set found a single most parsimonious tree of 1229 steps, with a topology nearly identical to that of 26S rDNA tree (Fig. 5). The only difference between the two trees involves the placement of *C. volkensii* within the cornelian cherries. In the combined tree, *C. volkensii* is placed as the sister of *C. sessilis*, whereas, in the 26S rDNA tree, the species is recognized as a distinct lineage sister to the remainder of the cornelian cherries. Significantly, higher bootstrap and decay values for all clades were obtained for the tree derived from the combined data (Fig. 5).

DISCUSSION

Phylogenetic potential of 26S rDNA sequences below the genus level -- The phylogenetic utility of 26S rDNA sequences in seed plants has been demonstrated only recently by a few studies of taxa at higher taxonomic levels (above the generic level) (e.g., Ro, Keener, and McPheron, 1997; Kuzoff et al., 1998; Stefanovic et al., 1998; Ashworth,

2000). Our phylogenetic study of *Cornus* using 26S rDNA sequences provides an example of phylogenetic utility of this gene at an intrageneric level. Although the rate of evolution of 26S rDNA is relatively conservative (comparable to *rbcL*), the gene contains several expansion segments that evolve faster than the conserved core regions (see Bult, Sweere, and Zimmer, 1995; Kuzoff et al., 1998). The variable rate of evolution in the expansion segments and the conserved core regions makes the gene useful for phylogenetic analyses at different taxonomic levels depending on the regions employed (see Kuzoff et al., 1998). The expansion segments can be used at lower taxonomic levels, whereas the conserved core regions are suitable at higher levels. In addition, the large size of the gene provides more variable characters for phylogenetic analysis than *rbcL*; the low sequence variation as a result of the conservative rate of the gene is well compensated by its large size. Although only 11.56% of the sites is variable, and 4.05% is phylogenetically informative in the 26S rDNA matrix of *Cornus*, the total number of variable sites from the entire sequence are 391, and the number of total informative sites are 137. These numbers are nearly two times higher than those for *rbcL* and *matK*, respectively, in the combined data matrix (see Results). Analyses of the 26S rDNA sequences of *Cornus* resulted in a completely resolved phylogeny of the genus and suggested relationships congruent with the cpDNA-based phylogeny (Figs. 3, 4). This suggests that the 26S rDNA sequences as a whole contain sufficient phylogenetic information at intrageneric level of *Cornus* and are useful for elucidating phylogenetic relationships within the genus.

To explore differential utilities of the expansion segments and conserved core regions of 26S rDNA sequences in *Cornus*, we conducted phylogenetic analyses of the sequences from the expansion segments and conserved core regions separately. The analyses of these

portions of 26S rDNA did not produce topologies that were as fully resolved. The results indicated that the analysis of the expansion segments alone produced a weakly supported phylogeny inconsistent with both the cpDNA-based phylogeny and the morphology-based phylogeny. In this phylogeny, the monophyly of the cornelian cherries was unsupported and species of the group appear in different clades. Similar results were obtained from the analysis of the conserved core regions. In the trees resulting from the analysis of the conserved core sequences, the monophyly of *Cornus* was even unsupported. These odd results may be explained as the effect of insufficient variable characters in either the expansion segments or the conserved core regions. Although the expansion segments contain a higher percentage of phylogenetically informative sites (9.38%) than that of the entire 26S rDNA (4.05%), the total number of phylogenetically informative sites is reduced to 97 bp. The conserved core regions contain only 40 potentially phylogenetically informative sites. Moreover, the expansion segments have a significantly higher average G+C content (68.5%) than that of the entire 26S rDNA (58.1%) and that of the conserved core regions (52.6%). The higher G+C content in the 26S rDNA expansion segments poses a potential problem for phylogenetic analysis if the algorithm used to construct a tree assumes equally abundant nucleotides (see Kuzoff et al., 1998).

Phylogenetic relationships within Cornus - - The phylogenetic tree derived from the analysis of 26S rDNA sequences of *Cornus* shows a topology identical to the tree derived from cpDNA data (Fig. 1; Xiang et al., 1996; Xiang, Soltis, and Soltis, 1998), although some nodes in the 26S rDNA tree are not strongly supported by bootstrap and decay analyses (Figs. 3, 4). This 26S rDNA tree is also congruent with the tree derived from analyses of the combined 26S rDNA and cpDNA data (Fig. 5). Thus all of the molecular data from both

nuclear and chloroplast genomes of *Cornus* are concordant. These data suggest that the genus diverged early into two large lineages: (1) the blue- or white-fruited group, and (2) the red-fruited group. The red-fruited group subsequently separated into the cornelian cherries and a clade bearing showy bracts, which then diverged into the big-bracted dogwoods and the dwarf dogwoods. Within the blue- or white-fruited group, *C. oblonga* was the first lineage to branch off. This phylogenetic pattern is consistent with the scheme proposed by Eyde (1988), but at odds with the morphology-based phylogeny (Murrell, 1993). Phylogenetic analysis of morphological characters by Murrell (1993) placed the cornelian cherries (rather than the dwarf dogwoods as in the molecular tree) as the sister of the big-bracted dogwoods. This relationship was supported by five inflorescence characters which were treated as independent characters, including protective bracts, precocious peduncle, inflorescence preformed in the previous fall and developed from a mixed bud, reduced primary inflorescence branches, and reduced secondary inflorescence branches. However, three of these characters (protective bracts, precocious peduncle, and reduced secondary inflorescence branches) are homoplasious on the morphological trees, where they are reversed one or two times in some other clades or terminal taxa. In addition, these five inflorescence characters may not be independent because these characters may be developmentally correlated. Thus, differences in inflorescence may be overweighted in the morphological analysis, resulting in an incongruence between the morphological and molecular phylogeny. Based on the molecular phylogeny derived from multiple molecular data sets, the synapomorphies used to unite the cornelian cherries and the big-bracted dogwoods appear to be plesiomorphic characters evolved in the ancestor of the red-fruited group, but were later lost in the dwarf dogwoods (see discussion in Xiang et al., 1996).

Alternatively, these features may have evolved independently in the big-bracted dogwoods and the cornelian cherries. However, these two hypotheses cannot be distinguished without fossils representing the ancestor of the red-fruited group.

The second major incongruence between the molecular and morphological phylogenies involves the placement of *C. oblonga* (subg. *Yinquania*). *Cornus oblonga* was placed as the basal group within *Cornus* in the morphological tree (Fig. 2; Murrell, 1993), whereas in the molecular phologeny, *C. oblonga* is a distinct lineage sister to the remainder of the blue- or white-fruited dogwoods, a relationship strongly supported by bootstrap and decay analyses (Figs. 3, 5). A single morphological character (displaced bracts, present in the remainder of the genus, absent in *C. oblonga*) separated *C. oblonga* and the remainder of the genus in the morphological tree (Murrell, 1993). Based on the molecular phylogenies, this character is better explained as a plesiomorphic condition, and the nondisplaced bract in *C. oblonga* is a derived state (perhaps a reversal). Several morphological, anatomical, biochemistry characters (blue fruits, lack of iridoids, crassinucellate ovule, open cyme with minute bracts) also support the placement of *C. oblonga* as member of the blue- or white-fruited group.

The tropical African species *C. volkensii* is the only dioecious species in *Cornus*. Due to its morphological uniqueness, the species was treated as a separate subgenus (subg. *Afrocrania* Harms) or as a distinct genus (*Afrocrania* Hutch.) (Hutchinson, 1942; Ferguson, 1966; also see Murrell, 1993; Xiang et al., 1993). Eyde (1988) considered *C. volkensii* as a member of the cornelian cherries, based mainly on similarities in fruit morphology and inflorescence type between *C. volkensii* and other cornelian cherries. Our 26S rDNA sequence data and the combined 26S rDNA-cpDNA data strongly supported *C. volkensii* being a member of the cornelian cherries (Figs. 3, 5) and recognized it as either the sister of

all of the other cornelian cherries (Fig. 3) or as the sister of *C. sessilis* (Fig. 5). However, these placements of *C. volkensii* are not highly supported by both bootstrap and decay analyses (Figs. 3, 5). Similar to the 26S rDNA analysis, cladistic analysis of morphological characters also suggested *C. volkensii* is sister to the rest of the cornelian cherries (Murrell, 1993; Fig. 3). No synapomorphic characters can be found at present the sister relationship between *C. volkensii* and *C. sessilis*. Further, Eyde (1988) also proposed that *C. volkensii* was the first branched lineage within the cornelian cherries; its divergence from the cornelian cherries might have occurred in the Paleocene or early Eocene based on the morphological characters and fossil evidence. Therefore, the phylogenetic relationship of *C. volkensii* within the cornelian cherries based on 26S rDNA sequence data is concordant with morphological data and agrees with Eyde's hypothesis.

The dwarf dogwoods are the only herbaceous members of the dogwood genus *Cornus*. This group comprises three species, *C. canadensis, C. suecica*, and *C. unalaschkensis*. Evidence from cytology, phytogeography, and morphology suggests that *C. unalaschkensis* may be an allotetroploid species derived from past hybridization between *C. canadensis* and *C. suecica* followed by chromosomal doubling (Love and Love, 1975; Bain and Denford, 1979; Murrell, 1994). The 26S rDNA sequence data suggested that *C. unalaschkensis* is more closely related to *C. canadensis* than it is to *C. suecica*. This result may indicate that, if *C. unalaschkensis* is indeed an allotetroploid, as has been hypothesized, the 26S rDNA in *C. unalaschkensis* has converted to the type of *C. canadensis*. However, more extensive analyses (e.g., analyses of both nuclear and cpDNA data including all three species with more extensive sampling) are needed to test the hypothesis and determine whether *C. unalaschkensis* is of hybrid origin.

In summary, the entire 26S rDNA sequence of *Cornus* show a low level of sequence divergence, but because of its great length, provides sufficient variable characters (compared to *rbcL* and *matK*) to resolve the phylogenetic relationships within *Cornus*. Analyses of 26S rDNA sequence data result in a phylogeny of *Cornus* that is congruent with that inferred from combined cpDNA data, but at odds with the phylogeny derived from morphological analyses. Combining data for phylogenetic analysis can minimize sampling error and maximize the explanatory power of the data if congruent hypotheses are generated from separate analyses (see review by Johnson and Soltis, 1998). This was also demonstrated in our analysis of the combined 26S rDNA and cpDNA data in *Cornus*, which not only suggested congruent phylogenetic relationships to those inferred from separate data analyses within *Cornus*, but also significantly increased supports for most of the clades recognized in the tree (compare Figs. 3 and 5).

LITERATURE CITED

ASHWORTH, VANESSA E. T. M. 2000. Phylogenetic relationship in Phoradendreae (Viscaceae) inferred from three regions of the nuclear ribosomal cistron. I. Major lineages and paraphyly of *Phoradendron*. *Systematic Botany* 25: 349-370.

BAIN, J. F., AND K. E. DENFORD. 1979. The herbaceous members of the genus *Cornus* in NW North America. *Botanical Notiser* 132: 121-129.

BALDWIN, B. G., M. J. SANDERSON, J. M. PORTER, M. F. WOJCIECHOWSKI, C. S. CAMPBELL, AND M. J. DONOGHUE. 1995. The ITS region of nuclear ribosomal

DNA: a valuable source of evidence on angiosperm phylogeny. *Annals of the Missouri Botanical Garden* 82: 247-277.

BREMER, K. 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. *Evolution* 42: 795-803.

BULT, C. J., J. A. SWEERE, AND E. A. ZIMMER. 1995. Cryptic sequence simplicity, nucleotide composition bias, and molecular coevolution in the large subunit of ribosomal DNA in plants: implications for phylogenetic analysis. *Annals of the Missouri Botanical Garden* 82: 235-246.

CLARK, G. B., B. W. TAGUE, V. C. WARE, AND S. A. GERBI. 1984. *Xenopus Laevis* 28S ribosomal RNA: a secondary structure and its evolutionary and functional implications. *Nucleic Acid Researches* 12: 6197-6220.

DOVER, G. A., AND R. B. FLAVELL. 1984. Molecular coevolution: DNA divergence and the maintenance of function. *Cell* 38: 622-623.

EERNISSE, D. J., AND A. G. KLUGE. 1993. Taxonomic congruence versus total evidence, and amniote phylogeny inferred from fossils, molecules, and morphology. *Molecular Biology and Evolution* 10: 1170-1195.

EYDE, R. H. 1987. The case for keeping *Cornus* in the broad Linnaean sense. *Systematic Botany* 12: 505-518.

EYDE, R. H. 1988. Comprehending *Cornus*: puzzles and progress in the systematics of dogwoods. *Botanical Review* 54: 233-351.

FELSENSTEIN, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39: 783-791.

FERGUSON, I. K. 1966. Notes on the nomenclature of *Cornus*. *Journal of the Arnold Arboretum* 47: 100-105.

FLAVELL, R. B. 1986. Structure and control of expression of ribosomal RNA genes. *Oxford Surv. Plant Molecular Cell Biology* 3: 252-274.

HARMS, H. 1898. Cornaceae. *In* A. Engler and K. Prantl (eds.), *Die naturalichen Pflanzenfamilien*, Teil III, Abteilung 8, 250-270. Leipzig: W. Engelmann.

HARDIN, J. W., AND Z. E. MURRELL. 1997. Foliar micromorphology of *Cornus*. *Journal of the Torrey Botanical Society* 124:124-139.

HASEGAWA, M., H. KISHINO, AND T. YANO. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* 21: 160-174.

HUTCHINSON, J. 1942. Neglected genetic characters in the family Cornaceae. *Annals of Botany* (London), series 2(6): 83-93.

JONHSON, L. A., AND D. E. SOLTIS. 1998. Assessing congruence: Empirical examples from molecular data. *In* D. E. Soltis, P. E. Soltis, and J. J. Doyle (eds.), *Molecular Systematics of Plants* II, 297-347. Chapman and Hall, New York, USA.

KUZOFF, R. K., J. A. SWEERE, D. E. SOLTIS, P. S. SOLTIS, AND E. A. ZIMMER. 1998. The phylogenetic potential of entire 26S rDNA sequences in plant. *Molecular Biology Evolution* 15: 251-263.

LOVE, A., AND D. LOVE. 1975. Cytotaxonomic Atlas of the Actic Flora. Vaduz: J. Cramer, Germany.

MISHLER, B. D., L. A. LEWIS, M. A. BUCHHEIM, K. S. RENZAGLIA, D. J. GARBARY, C. F. DELWIICHE, F. W. ZECHMAN, T. S. KANTZ, AND R. L.

CHAPMAN. 1994. Phylogenetic relationships of the "green algae" and "bryophytes." *Annals of the Missouri Botanical Garden* 81: 451-483.

MURRELL, Z. E. 1993. Phylogenetic relationship in *Cornus* (Cornaceae). *Systematic Botany* 18: 469-495.

MURRELL, Z. E. 1994. Dwarf dogwoods: intermediacy and the morphological landscape. *Systematic Botany* 19: 539-556.

NAKAI, T. 1909. Cornaceae in Japan. *Botanical Magzine Tokyo* 23: 35-45.

POJARKOVA, A. I. 1950. K voprosu o sistematicheskikh otnosheniyakh vnutri linneevskogo roda *Cornus* L. Bot. *Mater. Gerb. Bot. Inst. Komarova Akad. Nauk.* SSSR 12: 164-180

RAFINESQUE, C. S. 1838. Alsograph. Am. Philadelphia: C. S. Rafinesque.

RO, K., C. S. KEENER, AND B. A. MCPHERON. 1997. Molecular phylogenetic study of the Ranunculaceae: utility of the nuclear 26S ribosomal DNA in inferring intrafamilial relationships. *Molecular Phylogenetics and Evolution* 8: 117-127.

SOLTIS, D. E., AND P. S. SOLTIS. 1997. Phylogenetic relationships among Saxifragaceae sensu lato: a comparison of topologies based in 18S rDNA and *rbc*L sequences. *American Journal of Botany* 84:504-522.

SOLTIS, D. E., AND P. S. SOLTIS. 1998. Choosing an approach and an appropriate gene for phylogenetic analysis. *In* D. E. Soltis, P. E. Soltis, and J. J. Doyle, (eds.), *Molecular Systematics of Plants* II, 1-42. Chapman and Hall, New York, USA.

SPACH, E. 1839. Les Cornaceae. *In Histoive Naturelle des Vegetaux*, vol. 8, 86-110. Paris: Librarie Encyclopedique de Roret.

STEFANOVIC, S., M. JAGER, J. DEUTSCH, J. BROUTIN, AND M. MASSELOT. 1998. Phylogenetic relationships of conifers inferred from partial 28S rDNA gene sequences. *American Journal Of Botany* 85: 688-697.

SUGIURA, M., Y. IIDA, K. OONO, AND F. TAKAIWA. 1985. The complete nucleotide sequence of a rice 25S rRNA gene. *Gene* 37: 255-259.

SWOFFORD, D. L. 2000. PAUP: phylogenetic analysis using parsimony, version 4.0b4. Sinauer, Sunderland, Massachusetts, USA.

VON BERCHTOLD, F., AND P. M. OPIZ. 1838. Cornaceae. *In Oekonomisch-technische Flora Bohmens*, Band 1, Teil 2, 167-180. Prag: T. Thabor.

WANGERIN, W. 1910. Cornaceae. *In* A. Engler (ed.), *Das Pflanzenreich*, series IV, family 229 (Heft 41). Leipzig: W. Engelmann.

XIANG, Q.-Y. 1987. A neglected character of *Cornus* L. s. l. with special reference to a new subgenus-*Sinocornus* Q. Y. Xiang. *Acta Phytotaxonomica Sinica* 25: 125-131.

XIANG, Q.-Y., D. E. SOLTIS, D. R. MORGAN, AND P. S. SOLTIS. 1993. Phylogenetic relationships of *Cornus* L. Sensu Lato and putative relatives inferred from *rbcL* sequence data. *Annals of the Missouri Botanical Garden* 80: 723-734.

XIANG, Q.-Y., S. J. BRUNSFELD, D. E. SOLTIS, AND P. S. SOLTIS. 1996. Phylogenetic relationship in *Cornus* based on chloroplast DNA restriction sites: implications for biogeography and character evolution. *Systematic Botany* 21: 515-534.

XIANG, Q.-Y., D. E. SOLTIS, AND P. S. SOLTIS. 1998. Phylogenetic relationships of Cornaceae and close relatives inferred from *matK* and *rbcL* sequences. *American Journal of Botany* 85: 285-297.

TABLE 1. Morphological characteristics of the subgenera of *Cornus*. All subgenera are woody and hermaphroditic with terminal inflorescence and opposite leaves except those indicated. Taxonomic treatment was synthesized from Ferguson (1966), Xiang (1987), Murrell (1993). Common names of subgenera are provided below the scientific names.

Subgroup	Fruits & Inflorescence	Size & Distribution
Subg. *Yinquania* (Zhu) Murrell (Oblong-blue-fruited dogwoods)	Blue, oblong fruits; open compound cymes; bracts minute.	1 spp., E. Asia
Subg. *Kraniopsis* Raf. (Blue- or white-fruited dogwoods)	Blue or white, and round fruits; open compound cymes; bracts minute.	~30 spp.; mostly E. Asia & N. Am., 2 or 3 Eur., 1 or 2 S. Am.
Subg. *Mesomora* Raf. (Alternate-leaved, blue-fruited dogwoods)	Blue fruits; alternate leaves; open compound cymes; bracts minute.	2 spp; E. Asia & E. N. Am.
Subg. *Afrocrania* Harms (African cornelian cherry)	Red fruits; umbellate cymes subtended by four nonshowy bracts; dioecious.	1 spp., Tropical Africa.
Subg. *Cornus* (Cornelian cherries)	Red fruits; umbellate cymes terminal subtended by four nonshowy bracts.	3 spp., E. Asia, W. N. Am., & Eur.
Subg. *Sinocornus* Q. Y. Xiang (Chinese cornelian cherry)	Red fruits; umbellate cymes axillary, subtended by four nonshowy bracts.	1 spp., China.
Subg. *Discocrania* (Mexican disciflorous dogwood)	Red fruits; capitulate cymes subtended by four early deciduous bracts.	1 or 2 spp., C. Am.

TABLE 1 continued

Subgroup	Fruits & Inflorescence	Size & Distribution
Subg. *Cynoxylon* Raf. (American big-bracted dogwoods)	Red fruits; capitulate cymes subtended by four large, showy bracts; fruit separate.	2 or 3 spp.; E. N. Am. & W. N.Am., ext. to Mexico.
Subg. *Syncarpea* (Nakai) Xiang. (Asian big-bracted dogwoods)	Red fruits; capitulate cymes subtended by four large, showy bracts; fruits fused.	4-12 spp; E. Asia.
Subg. *Arctocrania* Endl. Ex Reichenbach. (Dwarf dogwoods)	Red fruit; minute cymes subtended by four small, showy bracts; herbaceous.	3 spp.; Circumboreal.

TABLE 2. Species sampled in the study of 26S rDNA sequencing of *Cornus*. Species of *Cornus* are listed according to classifications of Ferguson (1966), Murrell (1993), and Xiang (1987).

Species	Sources and location of vouchers	GenBank accession no.
Cornus L.		
Subgen. *Yinquania*		
C. oblonga Wall.	Sun, s.n., Bot. Gard. Kunming, China	AF297539
Subgen. *Mesomora*		
C. controversa Hemsl.	Arnold Arboretum 20458, WS.	AF297541
Subgen. *Kraniopsis*		
C. racemosa Lam.	Xiang et al. 157, WS.	AF297538
C. walteri Wangerin	Arnold Arboretum 414-67-1, WS.	AF297540
Subgen. *Afrocrania*		
C. volkensii Harms	Knox 2528, Africa.	AF297542
Subgen. *Cornus*		
C. mas L.	Arnold Arboretum 577-51-A, WS.	AF297535
C. officinalis Seib. et Zucc.	Boufford et al. 26065, GH.	AF297536
C. sessilis Torr. Ex Durand	Terry M. Hardig, California 1994.	AF297537
Subgen. *Arctocrania*		
C. canadensis L.	Xiang et al. 198, WS.	AF297530
C. unalaschkensis Ledeb.	Xiang 210, WS.	AF297534
C. suecica L.	Chris Brochmann, 94-388, Norway.	AF297531

TABLE 2 continued

Species	Sources and location of vouchers	GenBank accession no.
Subgen. *Cynoxylon*		
C. florida L.	Xiang 250, WS.	AF297532
Subgen. *Syncarpea*		
C. kousa Hance	Xiang 310, Ohio State University compus.	AF297533
Outgroups		
Nyssa ogeche Marsh.	U.S. National Arbortum, s.n.	AF297545
Alangium platanifolium (Sieb. & Zucc.) Harms	Soltis 2543, Japan	AF297544
Schizophragma hydrangeoides Sieb. et Zucc.	Soltis 2516, Japan	AF297543

TABLE 3. Twenty insertions and deletions (lettered A-T) inferred from 26S rDNA sequences of *Cornus* and outgroups. An asterisk "*" indicates indels that are potentially phylogenetically informative within *Cornus* (the indels can be present on one or more than one lineages).

Indel	Sequence involving the indels	Base position in the aligned sequences	Species with the indel sequence present
A*	C	8	*C. walteri, C. controversa, Nyssa*
B	C	432	*C. volkensii, Nyssa*
C	C/T	468	*C. canadensis, C. suecica, C. florida, C. kousa, C. unalaschkensis, C. mas, C. officinalis, C. sessilis, C. racemosa, C. oblonga, C. walteri, C. controversa, C. volkensii, Nyssa*
D*	C/G	591	*C. kousa, C. mas, C. officinalis, C. sessilis, C. racemosa, C. oblonga, C. walteri, C. controversa, C. volkensii*
E*	G	600	*C. canadensis, C. suecica, C. unalaschkensis*
F	G	606	*Nyssa*
G*	T	638	*C. canadensis, C. suecica, C. unalaschkensis*
H	CCCC	751-754	*C. volkensii*
I	GGG	761-763	*C. volkensii*
J	RK	1002-1003	*Alangium, Schizophragma, Nyssa*
K	C/T	2083	*Schizophragma, Nyssa*

TABLE 3 continued

L	AGG	2095-2097	*Schizophragma*
Indel	Sequence involving the indels	Base position in the aligned sequences	Species with the indel sequence present
M*	T	2156	*C. unalaschkensis, C. mas, C. controversa, Schizophragma*
N*	G	2209	*C. kousa, C. sessilis, C. oblonga, C. walteri, C. controversa, C. volkensii, Alangium, Schizophragma, Nyssa*
O*	A	2238	*C. sessilis, C. volkensii*
P*	C	2596	*C. sessilis, C. volkensii*
Q	CGC	3182-3184	*Alangium*
R	C	3257	*C. volkensii, Alangium, Schizophragma, Nyssa*
S	GGTGC	3270-3274	*Alangium*
T*	T	3358	*C. suecica, C. florida, C. recemosa, C. walteri, C. controversa, C. volkensii*

TABLE 4. Locations and lengths of the 12 expansion segments (D1-D12) in 26S rDNA sequence of *Cornus* L. Location positions are expressed with reference to the coordinates for expansion segment positions in the sequence of *Oryza sativa* (Kuzoff, et al. 1998). All sequences are written 5' → 3'.

Expansion segment	Position	Length	Sequence before ES in *Cornus* and outgroups	Sequence after ES in *Cornus* and outgroups
D1	113-261	149	GAANAGCCCA	ACGAGTCGGG
D2	422-652	231	GGGAGGGAAG	GCCCGTYTTG
D3	694-809	116	ACATGTGTGC	AGCATGCCTG
D4	1004-1010	7	GCTGGAGCCC	TTMTATCGGG
D5	1091-1129	39	ATAGGTAGGA	AGCTCCAAGT
D6	1162-1188	27	GTAAGCAGAA	GGKTTACCGT
D7a	1562-1605	44	TCGATCCTAA	AAAGGGAATC
D7b	1648-1673	26	ACGTGRCGGC	ACGTYGGCGG
D8	1970-2115	146	GCTCTGAGGG	CARCTGACTC
D9	2490-2515	26	GGATAAGTGG	CCACTACTTT
D10	2541-2619	79	TTATTTTACT	GACATTGTCA
D11	3021-3024	4	CCCTACTGAT	GTGYCGCAAT
D12	3179-3318	140	AGYSACGCAT	AGAATCCTTT

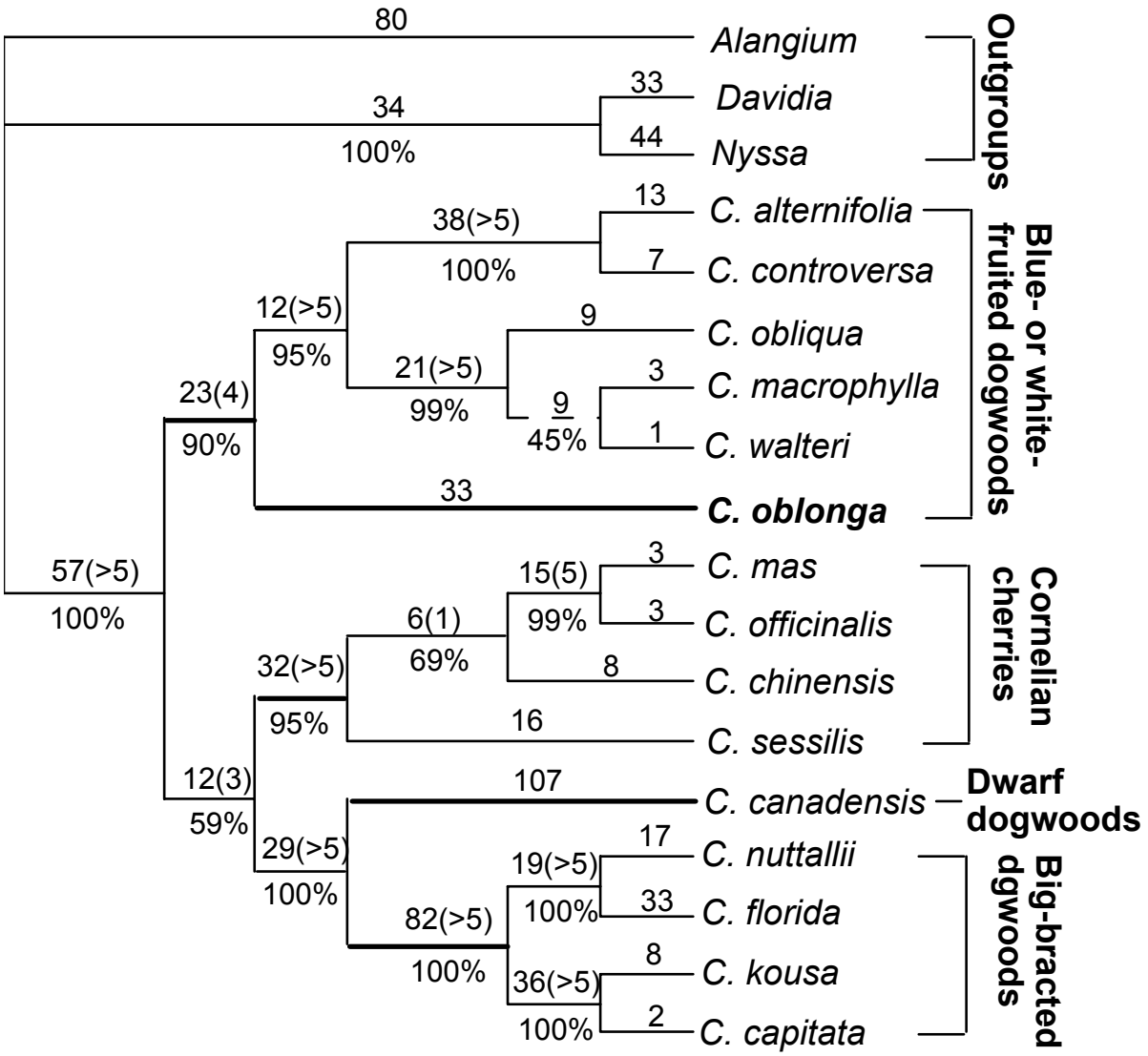


Fig. 1. One of the most parsimonious trees resulting from phylogenetic analyses of combined data set of *rbcL* and *matK* sequences and cpDNA restriction site data for *Cornus* (length = 845, CI = 0.707, excluding uninformative characters, and RI = 0.823) (modified from Xiang, Soltis, and Soltis, 1998). Base substitutions are indicated above branches; bootstrap values are indicated below branches; decay values are indicated by numbers in parentheses.

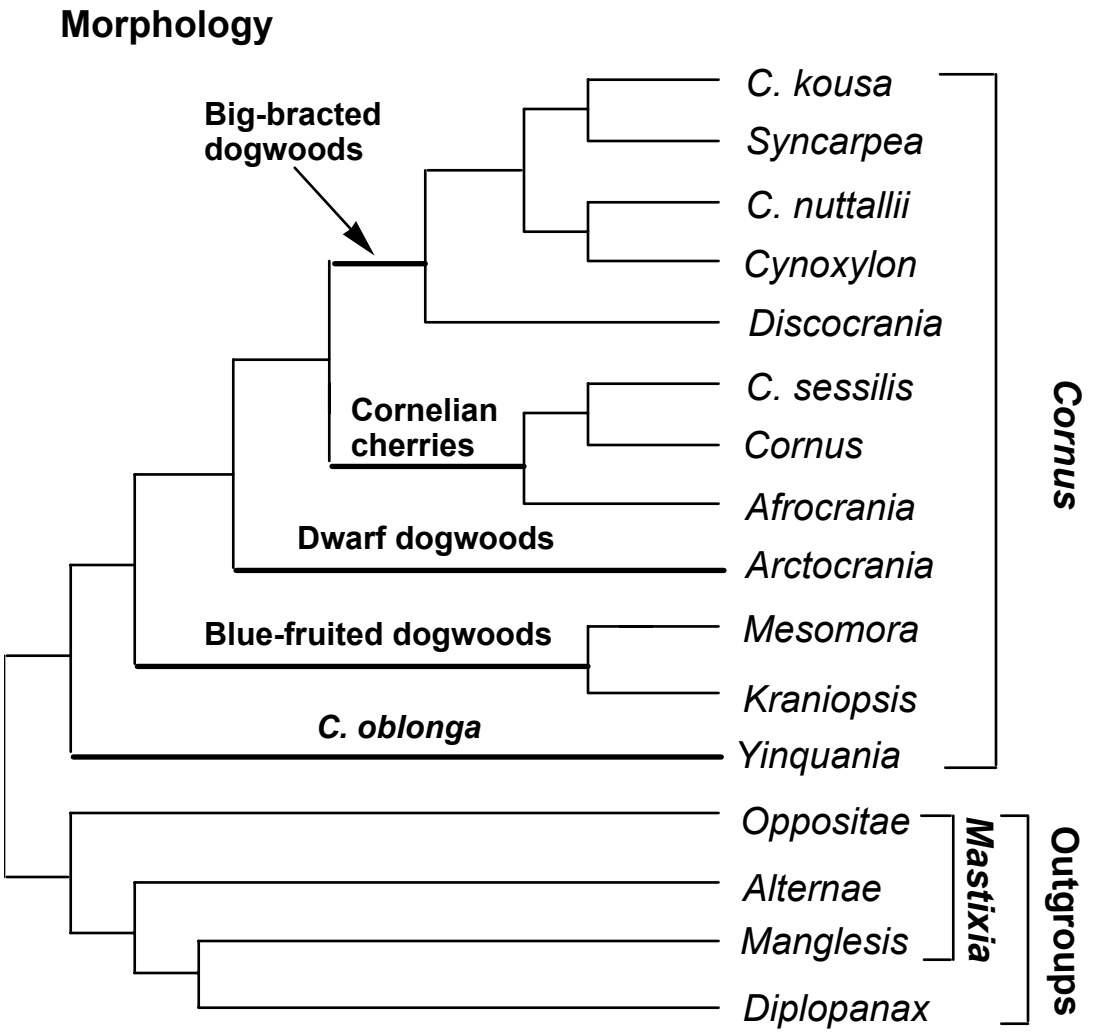


Fig. 2. The phylogenetic tree derived from cladistic analysis of 28 morphological, anatomical, chemical, and cytological characters modified from Murrell (1993).

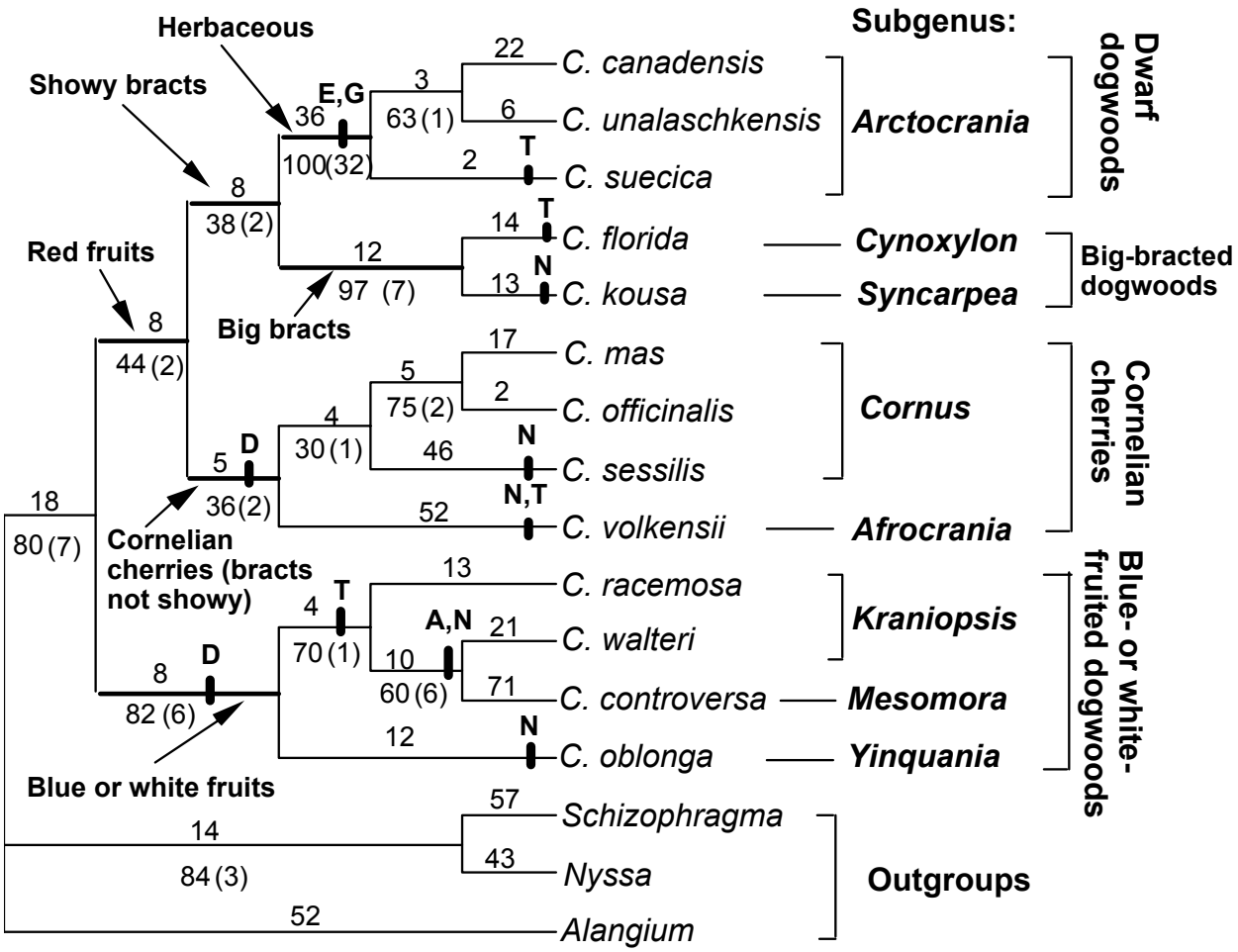


Fig. 3. The single most parsimonious tree resulting from analysis of the entire 26S rDNA sequences (tree length = 577 steps, CI = 0.781 excluding uninformative characters, RI = 0.619). Base substitutions are indicated above branches; bootstrap values are indicated below branches; decay values are indicated by numbers in parentheses. Indels were coded as missing. Six of phylogenetically informative indels (A, D, E, G, N, T) within *Cornus* are mapped on the branches. Sequence information regarding these indels is shown in Table 3.

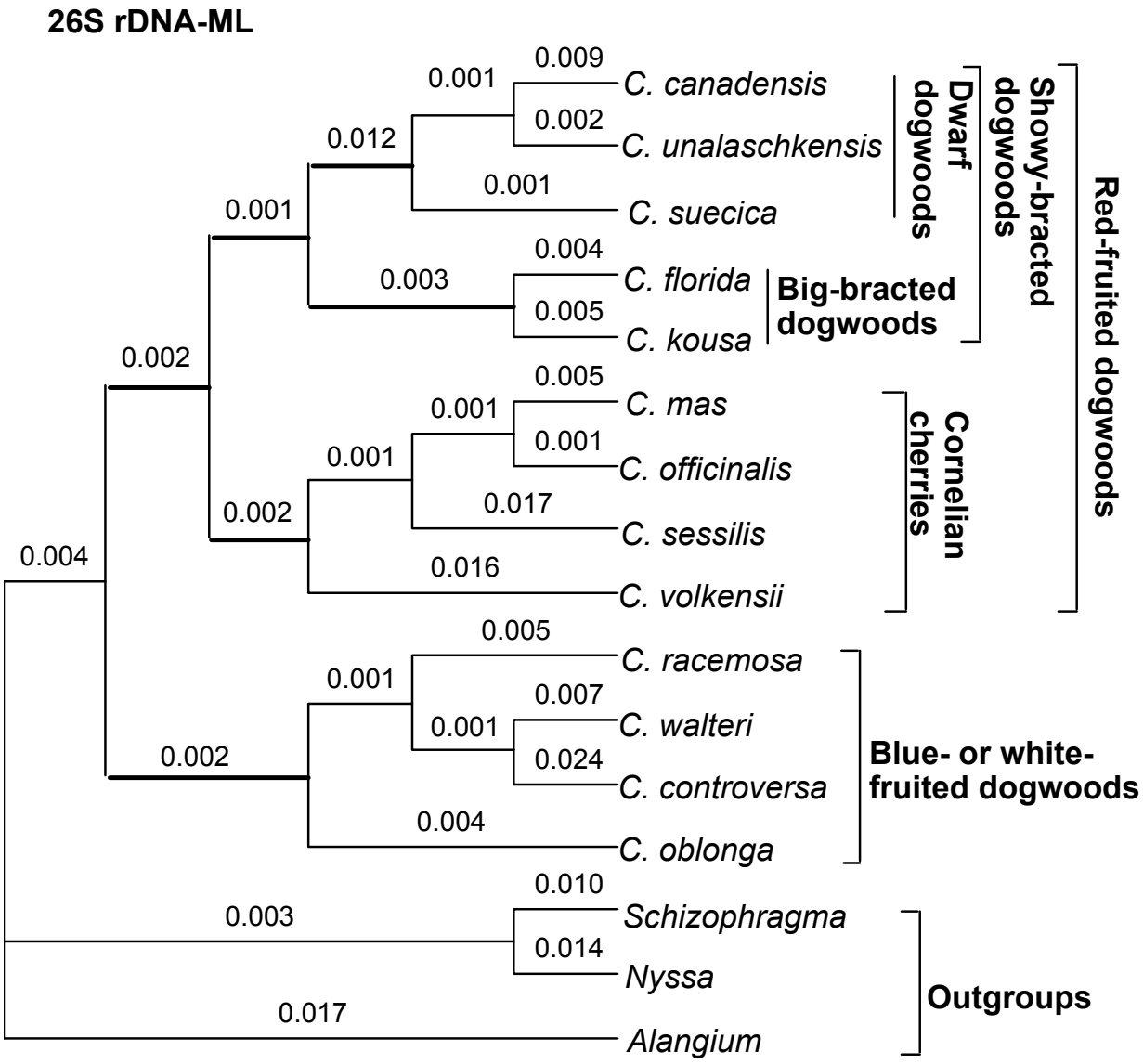


Fig. 4. The maximum likelihood tree resulting from analysis of the 26S rDNA sequences using heuristic searches with random taxon addition of five replicates, empirical base frequency, HKY model requested two parameter model variant for unequal base frequency, ti/tv = 2 [kappa = 2.402791], and equal rates for all sites; branch lengths are indicated above the nodes (-Ln likelihood = 8531.9677).

Combined 26S rDNA- *matK-rbcL*-cpDNA R.S.-Parsimony

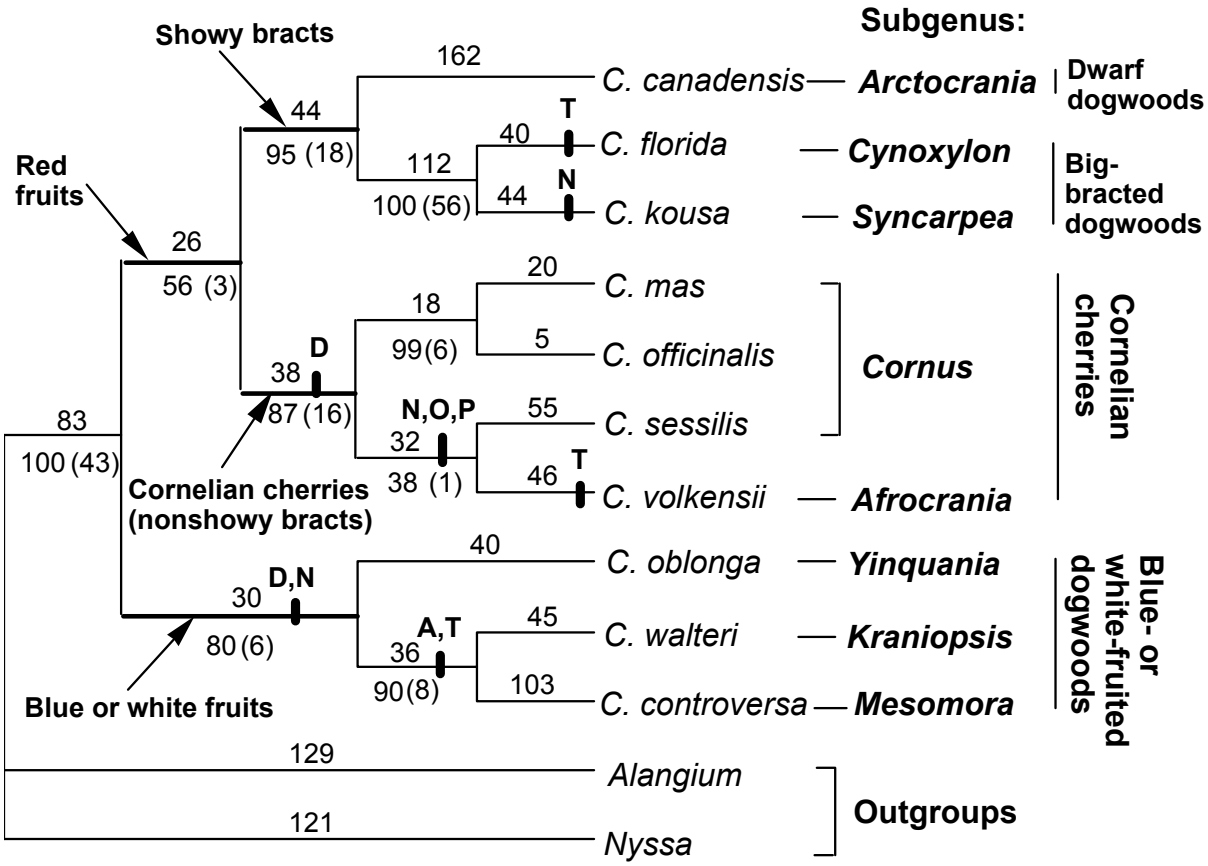


Fig. 5. The single most parsimonious tree resulting from analysis of combined 26S rDNA sequences and cpDNA data (*matK-rbcL* sequences and restriction sites) (tree length =1229 steps, CI = 0.830 excluding uninformative characters, RI = 0.627). Base substitutions are indicated above branches; bootstrap values are indicated below branches; decay values are indicated by numbers in parentheses. Six phylogenetically informative indels (A, D, N, O, P, T) from 26S rDNA sequences are mapped on the branches. Sequence information regarding these indels is shown in Table 3.

Chapter III

Heterogeneous Evolution of the *myc*-like Anthocyanin Regulatory Gene in *Cornus* (Cornaceae)

Chuanzhu Fan[1, 3], Michael D. Purugganan[2], David T. Thomas[1], Brian M. Wiegmann[3], and (Jenny) Qiu-Yun Xiang[1]

[1]Department of Botany, North Carolina State University, Raleigh, NC 27695-7612;

[2]Department of Genetics, North Carolina State University, Raleigh, NC 27695-7614

[3]Department of Entomology, North Carolina State University, Raleigh, NC 27695-7613

[3]Corresponding author:

Chuanzhu Fan, Department of Botany, North Carolina State University, Raleigh, NC 27695-7612.

Phone number: 919-515-3345

Fax number: 919-515-3436

Email: cfan3@unity.ncsu.edu

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

Manuscript in preparation: Follow the format of '*Molecular Biology and Evolution*'

Chuanzhu did most of the work reported in this chapter. David Thomas completed partial sequence of *C. chinensis* and *C. eydeana*. Drs. Michael D. Purugganan, Brian M. Wiegmann, and (Jenny) Qiu-Yun Xiang provided scientific advice and guidance.

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 and 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. Among the four functional domains, the interaction domain, involving an important function, evolves under the strongest evolutionary constraint, whereas the acidic domain is the most rapidly evolving region. The bHLH domain also evolves faster than the interaction domain, but the key residues are conserved, and most changes occur at loop region and among amino acids with similar chemical features. The elevated rate of evolution in both bHLH domain and acidic domain may contribute to the change of transactivation activity, which may alter the rates of transcription of anthocyanin synthetic genes.

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 of plant species (Mol, Grotewold, and Koes 1998). Studies have indicated that these plant pigments assist in pollinator attraction, fruit dispersal, pollen viability, plant disease defense, and UV protection (Epperson and Clegg 1987; Ludwig and Wessler 1990; Stapleton 1992; Durbin, McCaig, and Clegg 2000). 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, McCaig, and Clegg 2000; Spelt et al. 2000) and the other class regulates the anthocyanin biosynthetic genes, a gene family present in diverse plant species (Goodrich, Carpenter, and Coen 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.

Genetic studies of mutations in *Zea mays* identified two families of regulatory genes control the transcription of all anthocyanin biosynthetic structural genes, the *R* and *c1* families. *R* family genes that encode *myc*-like proteins which contain a basic helix-loop-helix (bHLH) motif found in other eukaryotic transcriptional factors (e.g. DePinho et al. 1987; Davis, Weintraub, and Lasser 1987). In maize, the *R* family includes of *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, Carpenter, and Coen 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 2001, 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 (Martin et al. 1991; Lloyd, Walbot, and Davis 1992; Consonni et al. 1992, 1993; Quattrocchio et al. 1993, 1998; Goldsbourough et al. 1996). 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, 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 and Wessler 1994; Purugganan 1998; 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, Robichaux, and Purugganan 2001), suggesting that the accelerated evolution rates in floral regulatory genes may be correlate 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 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 phylogenetic analyses (Xiang, Soltis, and Soltis 1998; Fan and Xiang 2001, 2003). Molecular data from these two chloroplast genes (*matK* and *rbcL*) and one nuclear gene (26S rDNA) are available for the dogwoods from previous studies (Xiang, Soltis, and Soltis 1998; Fan and Xiang 2001), 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.

Materials and Methods

Identification and Characterization of Gene 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-CAATGGAGYTATRTYTTHTGGTC and reverse primer R2-TCRGTRAGRTCTTCWGGDGATAATGC). The primer F1 is located at the beginning of 5'-end of the interaction domain (exon 1), and R2 is at the middle of interaction domain (exon 2). The relative positions of 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 a 800-bp length fragment. Cloning of the PCR products revealed two types (designated as Type 'A' and Type 'B') of sequences; both are highly similar to the anthocyanin regulatory gene in *Petunia*, *Perilla*, and *Antirrhinum majus* based on Blast search 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 from known sequences and locus specific/degenerated reverse primers from Genbank alignment sequences. 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.

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*) were 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 base pairs 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 Table 1. 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, Soltis, and Soltis (1998). PCR reactions were performed using different combinations of the forward and reverse primers described as above. PCR reactions contained the following: 5 μL of 10x Mg^{2+} free buffer, 6 μL of 25mmol/L $MgCl_2$, 6-10 μL of 2.5mmol/L dNTPs, 0.5μL of 20μmol/L forward primer, 0.5μL of 20μmol/L reverse primer, 5μL of DMSO (dimethyl sulfoxide), 1-5 μL of BSA (Bovine serum albumin, 10mg/ml), 0.3μL of *Taq* polymerase (Promega), 5-10 μL of 20ng/μL total DNA extract, and calibrated to final 50μL using sterile deionized water. In order to avoid non-specific primer annealing and increase yield of PCR products, a hot-start (six minutes 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 seconds for one cycle; (2) 30-40 cycles of 94°C for 45 sec, 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 (Table 2). 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 to competent *E. coli* cells using TOPO TA cloning techniques (Invitrogen Life Technologies, Carlsbad, California, USA 92008). The growing colonies were screened for positive transformants using PCR amplification by T3 and T7 primers located on 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, Wisconsin 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, California, USA). Cycle-sequencing reactions (10 μL) were prepared by combining 2 μL terminator ready reaction mix, 2 μL sequencing buffer (200 mmol/L Tris-pH8.0, 5mmol/L $MgCl_2$), 0.6 μL primer (5 μmol/L), 2 μL or 4 μL of 200 ng/μL cleaned PCR products or plasmid DNA, 0.5 μL DMSO, and 2.9 μL (for PCR product reactions) or 0.9μL (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 96ºC for 30 sec, 50ºC for 15 sec, and 60ºC for 4 min. Products of cycle-sequencing were cleaned using ethanol/sodium acetate precipitation (ABI

applied Biosystems, Foster City, California 94404, USA) with an additional 95% ethanol wash. The cleaned sequencing products were analyzed on an ABI-377 automated sequencer (Applied Biosystems, Foster City, California 94404, USA). The sequence chromatogram output files for all samples were checked and edited base by base manually before being aligned.

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 3). These included three species of dwarf dogwoods (*Cornus* subgen. *Arctocrania*), *C. canadensis*, *C. suecica*, and *C. unalaschkensis* and *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; Xiang et al. 1993; Xiang, Soltis, and Soltis 1998; Fan and Xiang 2001).

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*, *Antirrhinum majus*, *Zea 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 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 on PAUP* 4.0b10 using heuristic searches with random taxon addition of 100 replicates. To evaluate clade support in both parsimony and ML analyses, 10000 replicates of bootstrap analysis (Felsenstein 1985) were performed using fast heuristic search and TBR branch-swapping.

Rate and Pattern of Gene Evolution

Analysis of synonymous (Ks) and nonsynonymous (Ka) substitution rate: To examine the pattern and rate of nucleotide substitutions in the coding region, pairwise synonymous (Ks) and nonsynonymous (Ka) nucleotide substitution rates in the coding regions were examined. Ks and Ka 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 Ka/Ks was calculated by dividing the nonsynonymous nucleotide substitutions rate with the synonymous nucleotide substitution rate. Ka and Ks 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 (Ka/Ks) was plotted against the rate of synonymous substitution (Ks) 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 ω = Ka/Ks) 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). This codon-substitution model is a tree-based model, which estimates the ratios of Ka/Ks for entire and each codon sites based on tree input. 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). The one-ratio model (M0) assumes the same value of ω for all codon sites (Goldman and Yang 1994). The neutral model (M1) estimates a proportion (p_0) of the sequences as conserved sites with ω = 0 and a proportion $p_1 = 1-p_0$ of neutral sites with $\omega_1 = 1$ (Nielsen and Yang 1998). The selection model (M2) adds an additional class of sites with frequency $p_2 = 1-p_0-p_1$ and ω_2 estimated from the data (Nielsen and Yang 1998). The discrete model (M3) uses an unconstrained discrete distribution to model heterogeneous ω ratios among sites. The beta model (M7) calculates the beta distribution $B(p, q)$ of sites using the algorithm of Majumder and Bhattacharjee (1973). The latter model does not detect positive selection sites (with ω >1). Beta & ω model (M8) adds one extra class of sites to the beta model (M7) with a proportion

of p_0 sites having ω drawn from the beta distribution B (p, q), and the remaining sites (p_1 = 1-p_0) having the same ratio ω_1. 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 υ equal to the difference in the number of parameters between two models.

Results

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 *Antirrhinum 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*. In *C. oblonga* the gene lacks intron 4 (Fig. 1, Table 4). 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 highly conserved 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 5). The sizes of introns are highly variable in these *Cornus* species.

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 base pairs 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. A 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 202.63 (Table 6).

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, 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; Fig. 2, 3). Relationships among the major subgroups within *Cornus* based on the entire protein are consistent estimated from previous studies (Fan and Xiang 2001).

Phylogenetic analyses of nucleotides from the entire coding region 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 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 and 26S rDNA data analyses (Fan and Xiang 2001, 2003). However, major dogwoods 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).

Pattern and Rates of Gene Evolution

Rates of Ka and Ks- 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 Ka/Ks across all exons is 0.407±0.040 (mean ± 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, 6). However, the Ka/Ks 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, 6). Furthermore, plots of Ka/Ks versus Ks indicate that the ratio of Ka/Ks is positively related to Ks in the acidic domain and C-terminal domain, but negatively related to Ks in the interaction and bHLH domains (Fig. 7). These data suggest that Ka increases at a higher rate than Ks in the C-terminal and acidic domains, but Ka increases at a lower rate than Ks in the interaction and bHLH domains, where there

may be greater functional constraint. The individual pairwise Ka and Ks values and individual ratio of Ka/Ks are listed in Tables 7-14.

Sites under diversifying selection -The results of parameter estimation for different models using PAML with codeml are displayed in Table 15. 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 15). The LRT statistic for comparison of the neutral (M1) and selection models (M2) rejecte M1 in favor of M2 (Table 16). 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 ω_2 = 0.17).

The one-ratio model (M0) is rejected by a big margin when compared with model 3 (discrete model) (Table 16). 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 positively selected sites with P (ω>1) >0.5, among them 27 sites with P (ω>1)>0.95, and 6 sites with P (ω>1)>0.99. Similarly, tests with model M8 (beta & ω) suggeste that 42%, or a total of 240 codon sites, are positively diversifying selected sites with P (ω>1) >0.5, among them 27 sites with P (ω>1)>0.95, and 5 sites with P (ω>1)>0.99.

Over half of 241 (240 for M8) sites with P (ω>1)>0.50 detected are located in the acidic domain (Table 17). Seventeen of 27 sites with P (ω>1)>0.99 are also found in the acidic domain (Table 17). Among six sites detected by M3 with P (ω>1)>0.99, four are located in the acidic domain, and one resides in the interaction and bHLH domain,

respectively (Tables 17, 18). The site 442 in the bHLH domain is not supported with P>0.99 by the M8 (Table 18).

Discussion

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, Donoghue, and Zhang 1997; Small et al. 1998; Sang and Zhang 1999; Small and Wendel 2000); *phyB*-Mathews and Sharrock 1996; Mathews and Donoghue 1999; Mathews, Tsai, and Kellogg 2000; *waxy*- *PgiC* -Gottlieb and Ford 1996; Mason-Gamer, Weil, and Kellogg 1998; *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 anthocyanin regulatory gene largely eliminate the need for cloning to obtain clean and unambiguous sequences. Recent studies of some nuclear regulatory genes, [e.g. two MADS-box genes: *Pistillata* (Bailey and Doyle 1999) and *Leafy* (Nishimoto, Ohnishi, Hasegawa 2003)] also demonstrate that the 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 an adequate number of phylogenetically informative sites, and nucleotide sequences of this gene are informative of phylogeny in the dicot genus *Cornus*. 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. family or above). Our phylogenetic analyses of protein sequences including *Cornus*, five additional dicot sequences, and four monocot sequences are highly congruent with the current view of angiosperm phylogeny (Fig 3, 4). Our data also show that alignable intron sequences among the species within subgroups of *Cornus* can be useful to elucidate 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).

Rates and Pattern of Gene Evolution

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

Since 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 (Purugganan 1998; Ting et al. 1998; Alvarez-Buylla et al. 2000; Olsen et al. 2002; Fridman and Zamir 2003). For example, a reduced level of protein sequence polymorphism was detected in some developmental genes in *Drosophila* and maize (Purugganan 2000) compared to regulatory genes. The data from maize indicated that the ratios of nonsynonymous and synonymous substitutions of regulatory genes are significantly higher than those of structural genes (Gaut and Doebley 1997; Purugganan 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, Wu, and Luo 1985; Nei 1987; Purugganan 1998). Congruent with previous findings (Graur 1985), the relationships between Ka/Ks and Ks in *Cornus* further indicated that the synonymous and nonsynonymous rates are significantly correlated, although in different fashions among domains (Fig. 7). 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% variable 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 (Murre, McCaw, and Baltimore 1989; Atchley and Fitch 1997; Atchley et al. 2000). The function of the interaction domain involves transcriptional activation (Goff, Cone, and Chandler 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 Ka and Ks in the interaction and bHLH domains are lower than those in other domains, and the Ka/Ks is negatively related to Ks (Fig. 7). The significantly lower ratio of Ka/Ks 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, Cone, and Chandler 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 acids (E) and arginine (R)], required for DNA binding, and hydrophobic residues [e.g. Leucine (L)] (Fig. 8) at helix regions requiring the formation of a dimer, conserved among *Cornus* and other dicots.

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*; Fig. 8). However, the swap of amino acids mostly involves residues with similar chemical structures and/or charges (Fig. 8), e.g. leucine (L)-valine (V), serine(S)-threonine (T), glutamic acid (E)-aspartic acid (D), 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, Terhalle, and Dress 1999). Previous studies show that chemically similar amino acids are known to be more interchangeable than chemically different ones due redundancy of the genetic code and the effects of purifying selection (Gojobori, Li, and Graur 1982).

Our data suggest that the acidic domain evolves most rapidly. This domain has the highest synonymous substitution rates and Ka/Ks (Figure 5 and 6). Applying codon-based model indicates that over half of the potentially positive selection sites are from the acidic domain (126 of 241 sites, 52.3%, Table 17). Furthermore, four of the six positive sites identified as positively selected with 99% probability also occur in acidic domain. Amino acid substitutions at these four sites indicate that substitutions at three of them (sites 312, 312, and 378) involve changes between polar uncharged amino acids and nonpolar amino acids (Table 18, Fig. 8). 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 *Delia* in *Antirrhinum majus* found that *Delia* 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 need to verify this hypothesis.

In conclusion, the *myc*-like anthocyanin regulatory gene may provide useful markers for phylogenetic analyses at different taxonomic levels depending on the data set considered (DNA or protein sequences) and regions applied (exons and introns). Variability in regulatory genes of plants was demonstrated using the example of the *myc*-like anthocyanin gene in *Cornus*. Mosaic and heterogeneous gene evolution among different domains and sites was also detected. Among the four functional domains, the interaction domain, which involved the most important functions, evolves under the strongest evolutionary constrains, whereas the acidic domain is the most rapidly evolving region. The bHLH domain is evolving at a relatively higher rate than expected, but the key residues are conserved, and changes mainly occur among the amino acid residues with similar chemical features that would not substantially affect the function of protein. The elevated rates shown for the bHLH and acidic domains may increase transactivation activity that can significantly alter the transcription of anthocyanin synthetic genes.

Acknowledgments

The authors thank the following people for providing different kinds of help: Brian Wiegmann for sharing the sequencing facility and insightful comments of the original manuscript; 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 Deep Gene Travel Award from NSF grant to C.F.

Literature cited

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

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

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

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

Bailey, C. D., and J. J. Doyle. 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., R. H. Robichaux, and M. D. Purugganan. 2001. Accelerated regulatory gene evolution in an adaptive radiation. Proc. Natl. Acad. Sci. USA 98: 10208-10213.

Bate, N. J., and S. J. Rothstein. 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., A. Viotti, S. L. Dellaporta, and C. Tonelli. 1992. cDNA nucleotide sequence of *Sn*, a regulatory gene in maize. Nucleic Acids Res. 20: 373.

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

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

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

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

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

Epperson, B., and M. T. Clegg. 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., and Q.-Y. Xiang. 2001. Phylogenetic relationships within *Cornus* (Cornaceae) based on 26S rDNA sequences. Am. J. Bot. 88: 1131-1138.

Fan, C., and Q.-Y. Xiang. 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., and D. Zamir. 2003. Functional divergence of a synthetic invertase gene family in tomato, potato, and *Arabidopsis*. Plant Physiol.131: 603-609.

Gaut, B. S., and J. F. Doebley. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. Proc. Natl. Acad. Sci. USA 94: 6809-6814.

Goff, S. A., K. C. Cone, and V. L. Chandler. 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., W.-H. Li, and D. Graur. 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. 18: 360-369.

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

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

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

Gong, Z., E. Yamagishi, M. Yamazaki, and K. Saito. 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., R. Carpenter, and E. S. Coen. 1992. A common gene regulates pigmentation pattern in diverse plant species. Cell 68: 955-964.

Gottlieb, L. D., and V. S. Ford. 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., and E. C. Cornish. 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., B. Anderson, and S. Wessler. 1996. Isolation and characterization of rice genes: evidence for distinct evolutionary paths in rice and maize. Genetics 142: 1021-1031.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. *Pp* 21-132 *in* H. N. Munro, ed. Mammalian protein metabolism. Academic, New York.

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

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

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

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

Li, W.–H., Wu. C.-I., and C.-C. Luo. 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., and R. F. Whittier. 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., N. Mitsukawa, T. Oosumi, and R. F. Whittier. 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., V. Walbot, and R. W. Davis. 1992. *Arabidopsis* and *Nicotiana* anthocyanin production activated by maize regulator *R* and *C1*. Science 258: 1773-1775.

Ludwig, S. R., L. F. Habera, S. L. Dellaporte, and S. R. Wessler. 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., and S. R. Wessler. 1990. Maize *R* gene family: tissue-specific helix-loop-helix proteins. Cell 62: 849-852.

Majumder, K. L., and G. P. Bhattacharjee. 1973. The incomplete beta integral (AS63). Appl. Stat. 22: 409-411.

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

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

Mathews, S., and R. A. Sharrock. 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. Boil. Evol. 13: 1145-1150.

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

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

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

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

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

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

Murre, C., G. Bain, M. A. van Dijk, I. Engel, B. A. Furnari, M. E. Massari, J. R. Mathews, M. W. Quong, R. R. Rivera, and M. H. Stuiver. 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., and T. Gojobori. 1986. Simple method for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3: 418-426.

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

Nishimoto Y., O. Ohnishi, and M. Hasegawa. 2003. Topological incongruence between nuclear and chloroplast DNA trees suggesting hybridization in the urophyllum group of the genus Fagopyrum (Polygonaceae). Genes. Genet. Syst. 78: 139-53.

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

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

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

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

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

Posada, D., and K. A. Crandall. 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., and S. R. Wessler. 1994. Molecular evolution of the plant *R* regulatory gene family. Genetics 138: 849-854.

Quattrocchio, F., J. F. Wing, H. T. C. Leppen, J. N. M. Mol, and R. E. Koes. 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., J. F. Wing, K. van der Woude, J. N. M. Mol, and R. Koes. 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., D. Turks, and V. L. Chandler. 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., M. J. Donoghue, and D. Zhang. 1997. Evolution of alcohol dehydrogenase genes in Peonies (*Paeonia*): phylogenetic relationships of putative nonhybrid species. Mol. Biol. Evol. 14: 994-1007.

Sang, T., and D. Zhang. 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., J. A. Ryburn, R. C. Cronn, T. Seelanan, and J. F. Wendel. 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., and J. F. Wendel. 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., and P. S. Soltis. 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., F. Quattrocchio, J. N. M. Mol, and R. Koes. 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., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 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., S.-C. Tsaur, M.-L. Wu, and C.-I Wu. 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 regulation. Stadler Symp. 7: 117-134. Columbia, Mo.: University of Missouri.

Xiang, Q.-Y., D. E. Soltis, D. R. Morgan, and P. S. Soltis. 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., D. E. Soltis, and P. S. Soltis. 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., R. Nielsen, N. Goldman, and A.-M. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155: 431-449.

Table 1. Locus specific primers used for amplifying and sequencing entire genomic sequences of the *myc*-like anthocyanin regulatory gene for *Cornus*

Name of primer	Length (bp)	Sequence (5'-3')	Tm (^{0}C)	Location
F0A	21	TCACTGAGTGGGTGTCTTAAG	55.0	5'-Flank
F0A2	23	AGTCAAACCCTTTGAGTTTCTTC	57.0	5'-Flank
F1A2	18	AGTTATGGCTAGTAGTGG	48.2	Exon I
F1AB	24	GGTCCATTTCWTCAAGMCAGCCAG	60.8	Exon I
F2A	22	TTTATGAGTCCCTTGYGGTCAC	58.5	Exon II
F2A1	25	GTTCAGGCGGTCGAATTCAATGCCG	64.2	Exon II
R2A	22	GTGACCRCAAGGGACTCATAAA	57.2	Exon II
R2A2	20	CCACTCCGTATCCGTGAGGT	66.1	Exon II
R2A3	19	GAACGACATGCAAACCAAG	54.5	Exon II
F3'	20	GGAGGWGTRRTTGAGCTSGG	57.8	Exon V
R3'	20	CCSAGCTCAAYYACWCCTCC	57.8	Exon V
R3A	20	AGAGCGACTGAATATTTTGC	51.9	Exon III
F3A	20	CAGACTGTGGTATGCTTTCC	57.3	Exon V
F3A2	20	TTGCCRGGAAGAACRTTAGC	56.6	Exon III
F4A	20	GCGATATTGCCATTTGTCTG	53.2	Intron IV
F4A2	21	TCCCAATAACAATTCGAGTGG	56.0	Exon VI
F4A3	20	TATGGTAGAAGGCTTAAATG	47.9	Exon VI
R4A	21	CATTTATGGAAGTAAGGTCCC	51.6	Exon VI
R4A1	21	CATCATCCATGAATTGCCAGC	67.4	Exon VI

F5A	20	TGATCCCTTCCACTAGCAAG	54.9	Exon VI
F5A2	22	TCGGTCCTTGGATCATTGATCC	57.9	Exon VI
F5A1	22	GGCACAACCGACCTTTTCTTAG	57.4	Exon VI
F6A	19	CTGACCTCGTTGGACCTTC	55.6	Exon VI
F7A	19	AGGGAGCGCTTGTTGCTTG	59.5	Exon VII
R7A	19	CAAGCAACAAGCGCTCCCT	59.5	Exon VII
F7A2	19	GAGCTGGAGATCAACCTCG	55.4	Exon VII
R7A2	22	GATTGAACGGAGTGAGAATCTA	53.1	Exon VII
F7A3	25	GTTCAATTGAAAGATAGCTCAACTG	61.8	Exon VII
F7A4f	26	ATAGCTCAACTGATAATGTAACTGTC	58.4	Exon VII
R8A	22	CCATTTATGGGACTTTCTTAGT	51.4	3'-Flank
R8A2	21	GGAGATTCCCAAGAATTTGTG	52.7	3'-Flank
R8A3	21	AAGTGCTTGTATGATCATCCC	53.8	Exon VIII
R9A	21	CTATCCACAAGAAACACYTGC	53.8	3'-Flank

*F: forward primers; R-reverse primers

Table 2. Specific primers and arbitrary primers used for TAIL-PCR

Name	Length (bp)	Sequences (5'-3')	Tm (^{0}C)	Location	Notes
AD1	15	NTCGASTWTSGWGTT	46.0	Unknown	Liu et al 1995
AD2	16	NGTCGASWGANAWGAA	46.8	Unknown	Liu et al 1995
AD3	16	WGTGNAGWANCANAGA	34.8	Unknown	Liu et al 1995
R1A1	25	CTTATGCCAAAAGACATG TTCTTGA	58.1	Intron 1	Used for primary reaction to amplify the 5'end and flanking region
R1A2	25	TTGAATTCTAACTTACATC TACATG	54.8	Intron 1	Used for secondary reaction to amplify the 5'end and flanking region
R1A3	20	CATGGCACCCAATTATTAT T	51.2	Intron 1	Used for tertiary reaction to amplify the 5'end and flanking region
F5A1	22	GGCACAACCGACCTTTTCT TA	57.4	Exon VI	Used for primary reaction to amplify the 3'end and flanking region
F5A2	22	TCGGTCCTTGGATCATTGA TCC	57.9	Exon VI	Used for secondary reaction to amplify the 3'end and flanking region
F5A	20	TGATCCCTTCCACTAGCAA G	54.9	Exon VI	Used for tertiary reaction to amplify the 3'end and flanking region

*F: forward primers; R-reverse primers

Table 3. Sampling information

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

Table 4. Length of exon and intron (bp) of the *myc*-like anthocyanin regulatory gene

Species	*C. canadensis*	*C. suecica*	*C. unala-schkensis*	*C. florida*	*C. capitata*	*C. oblonga*	*C. alternifolia*	*C. chinensis*	*C. eydeana*	*Petunia JAF13*
Exon I	128	128	128	128	128	128	128	128	128	125
Exon II	255	255	255	255	255	255	255	255	255	278
Exon III	97	97	97	97	97	97	97	97	97	98
Exon IV	15	15	15	15	15	15	15	15	15	15
Exon V	57	57	57	57	57	57	57	57	57	57
Exon VI	813	825	828	825	825	840	840	825	825	837
Exon VII	426	426	426	444	444	444	444	444	444	417
Exon VIII	72	72	72	72	72	69	72	72	72	72
Exons sum	1863	1875	1878	1893	1893	1905	1908	1893	1893	1899
Intron 1	555	553	553	553	553	475	~555	555	526	625
Intron 2	531	530	530	517	522	475	547	451	486	82
Intron 3	199	199	199	200	184	180	240	206	230	487
Intron 4	109	109	109	118	118	0	118	114	114	221
Intron 5	102	102	102	109	109	110	110	111	111	102
Intron 6	116	116	116	103	103	110	111	104	104	85
Intron 7	255	255	255	191	324	247	324	205	229	186
Intron sum	1867	1864	1867	1791	1913	1597	2005	1746	1800	1788
Total	3730	3739	3745	3690	3806	3502	3913	3639	3693	3687
5'-flank	132	132	132	132	66	66	66	66	66	759
3'-flank	236	236	236	450	94	112	94	104	101	188

Entire sequence of *JAF*-13 was provided by Francesca Quattrocchio.

Table 5. 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	CTTGATSY DGMY
3	VI	Acidic	877-879	*C. oblonga, C. alternifolia*	3	ATT
4	VI	Acidic	958-969	*C. oblonga, C. alternifolia*	12	ATTGGTGG CTCT
5	VII	bHLH/C-terminal	1474-1491	*C. florida, C. capitata, C. oblonga, C. alternifolia, C. chinensis, C. eydeana*	18	TGCARGG AGSWRRC ARAK
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

Table 6. Absolute pair-wise distance matrix among nine *Cornus* species (the number of nucleotide differences)

Species	*C.cana-densis*	*C. unala-schkensis*	*C. suecica*	*C. florida*	*C. capitata*	*C. oblonga*	*C. alter-nifolia*	*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							

Table 7. Pairwise nonsynonymous (Ka) and synonymous (Ks) substitution rates (the number of Ka and Ks per site) between 11 *Cornus* samples for all exon regions. Estimated using Nei-Gojobori method with Jukes-Cantor model; number of sites=616.

	C. canadensis	*C. unalaschkensis*	*C. suecica-1*	*C. suecica-2*	*C. florida-1*	*C. florida-2*	*C. capitata*	*C. oblonga*	*C. alternifolia*	*C. chinensis*	*C. eydeana*
C. canadensis		0.0063	0.0230	0.0251	0.0765	0.0804	0.0700	0.1348	0.1410	0.0761	0.0909
C. unalaschkensis	0.0073		0.0194	0.0201	0.0761	0.0832	0.0726	0.1370	0.1431	0.0785	0.0917
C. suecica-1	0.0260	0.0286		0.0087	0.0831	0.0875	0.0805	0.1430	0.1482	0.0852	0.0985
C. suecica-2	0.0210	0.0235	0.0061		0.0838	0.0909	0.0828	0.1455	0.1511	0.0875	0.1008
C. florida-1	0.1918	0.1932	0.1932	0.1935		0.0077	0.0227	0.1139	0.1198	0.0622	0.0648
C. florida-2	0.2005	0.2019	0.2003	0.2022	0.0221		0.0219	0.1107	0.1150	0.0591	0.0617
C. capitata	0.1655	0.1670	0.1664	0.1667	0.0550	0.0579		0.1143	0.1195	0.0639	0.0666
C. oblonga	0.3627	0.3566	0.3520	0.3565	0.3000	0.2922	0.2911		0.0274	0.1115	0.1147
C. alternifolia	0.3653	0.3592	0.3552	0.3539	0.3030	0.2981	0.3042	0.0742		0.1183	0.1240
C. chinensis	0.2121	0.2152	0.2130	0.2133	0.1968	0.1903	0.1802	0.2781	0.2700		0.0266
C. eydeana	0.2191	0.2222	0.2200	0.2204	0.1800	0.1737	0.1638	0.2581	0.2672	0.0515	

Ka: above diagonal; Ks: below diagonal; *C. suecica*-1 from Alaska, USA; *C. suecica*-2 from Norway. *C. florida*-1 as voucher of 02-16 (Xiang 2002) from Veracruz, Mexico; *C. florida*-2 as voucher of 02-36 (Fan 2002) from North Carolina, USA.

Shade: Ks values below 0.05 are not used to calculate the ratio of Ka/Ks.

Table 8. Pairwise nonsynonymous (Ka) and synonymous (Ks) substitution rates (the number of Ka and Ks per site) between 11 *Cornus* samples from the interaction domain. Estimated using Nei-Gojobori method with Jukes-Cantor model; number of sites =193.

	C. canadensis	*C. unalaschkensis*	*C. suecica-1*	*C. suecica-2*	*C. florida-1*	*C. florida-2*	*C. capitata*	*C. oblonga*	*C. alternifolia*	*C. chinensis*	*C. eydeana*
C. canadensis		0.0090	0.0113	0.0170	0.0181	0.0205	0.0251	0.0228	0.0205	0.0381	0.0357
C. unalaschkensis	0.0076		0.0068	0.0079	0.0136	0.0251	0.0298	0.0275	0.0251	0.0429	0.0405
C. suecica-1	0.0076	0.0000		0.0056	0.0159	0.0274	0.0325	0.0302	0.0274	0.0456	0.0432
C. suecica-2	0.0115	0.0038	0.0038		0.0170	0.0332	0.0383	0.0360	0.0321	0.0516	0.0491
C. florida-1	0.0711	0.0626	0.0628	0.0673		0.0159	0.0209	0.0232	0.0205	0.0362	0.0338
C. florida-2	0.0873	0.0787	0.0790	0.0836	0.0463		0.0090	0.0068	0.0045	0.0171	0.0148
C. capitata	0.1045	0.0956	0.0945	0.0993	0.0450	0.0461		0.0159	0.0136	0.0287	0.0264
C. oblonga	0.1858	0.1756	0.1747	0.1805	0.1934	0.1656	0.1846		0.0068	0.0218	0.0194
C. alternifolia	0.1955	0.1850	0.1858	0.1773	0.1950	0.1749	0.2137	0.1561		0.0194	0.0171
C. chinensis	0.1698	0.1599	0.1590	0.1645	0.1772	0.1685	0.1780	0.1235	0.1323		0.0114
C. eydeana	0.1521	0.1424	0.1415	0.1468	0.1594	0.1509	0.1602	0.1240	0.1506	0.0298	

Ka: above diagonal; Ks: below diagonal; *C. suecica*-1 from Alaska, USA; *C. suecica*-2 from Norway. *C. florida*-1 as voucher of 02-16 (Xiang 2002) from Veracruz, Mexico; *C. florida*-2 as voucher of 02-36 (Fan 2002) from North Carolina, USA.

Shade: Ks values below 0.05 are not used to calculate the ratio of Ka/Ks.

Table 9. Pairwise nonsynonymous (Ka) and synonymous (Ks) substitution rates (the number of Ka and Ks per site) between 11 *Cornus* from the acidic domain. Estimated using Nei-Gojobori method with Jukes-Cantor model; number of sites =231.

	C. canadensis	*C. unalaschkensis*	*C. suecica-1*	*C. suecica-2*	*C. florida-1*	*C. florida-2*	*C. capitata*	*C. oblonga*	*C. alternifolia*	*C. chinensis*	*C. eydeana*
C. canadensis		0.0055	0.0330	0.0301	0.1153	0.1196	0.0994	0.2381	0.2311	0.1346	0.1545
C. unalaschkensis	0.0067		0.0310	0.0281	0.1184	0.1228	0.1024	0.2391	0.2321	0.1334	0.1533
C. suecica-1	0.0237	0.0306		0.0149	0.1238	0.1206	0.1098	0.2404	0.2308	0.1367	0.1566
C. suecica-2	0.0134	0.0203	0.0067		0.1205	0.1205	0.1066	0.2364	0.2294	0.1333	0.1532
C. florida-1	0.2887	0.2946	0.2609	0.2662		0.0037	0.0358	0.1778	0.1718	0.0632	0.0824
C. florida-2	0.2983	0.3043	0.2655	0.2756	0.0067		0.0397	0.1825	0.1741	0.0672	0.0865
C. capitata	0.1923	0.1972	0.1676	0.1721	0.0923	0.0998		0.1960	0.1874	0.0825	0.1021
C. oblonga	0.4657	0.4737	0.4316	0.4385	0.3607	0.3712	0.3749		0.0447	0.1732	0.1929
C. alternifolia	0.4641	0.4721	0.4299	0.4369	0.3571	0.3677	0.3713	0.0235		0.1718	0.1963
C. chinensis	0.3110	0.3172	0.2824	0.2879	0.2169	0.2258	0.1946	0.3098	0.3062		0.0309
C. eydeana	0.3330	0.3395	0.3035	0.3092	0.2406	0.2498	0.2176	0.3368	0.3332	0.0312	

Ka: above diagonal; Ks: below diagonal; *C. suecica*-1 from Alaska, USA; *C. suecica*-2 from Norway. *C. florida*-1 as voucher of 02-16 (Xiang 2002) from Veracruz, Mexico; *C. florida*-2 as voucher of 02-36 (Fan 2002) from North Carolina, USA.

Shade: Ks values below 0.05 are not used to calculate the ratio of Ka/Ks.

Table 10. Pairwise nonsynonymous (Ka) and synonymous (Ks) substitution rates (the number of Ka and Ks per site) between 11 *Cornus* samples from the bHLH domain. Estimated using Nei-Gojobori method with Jukes-Cantor model; number of sites = 56.

	C. canadensis	*C. unalaschkensis*	*C. suecica-1*	*C. suecica-2*	*C. florida-1*	*C. florida-2*	*C. capitata*	*C. oblonga*	*C. alternifolia*	*C. chinensis*	*C. eydeana*
C. canadensis		0.0078	0.0400	0.0400	0.0827	0.0917	0.0738	0.2013	0.2376	0.1093	0.1045
C. unalaschkensis	0.0256		0.0318	0.0318	0.0913	0.1005	0.0823	0.2114	0.2481	0.1137	0.1088
C. suecica-1	0.0801	0.1093		0.0000	0.1273	0.1370	0.1178	0.2483	0.2867	0.1507	0.1455
C. suecica-2	0.0524	0.0804	0.0260		0.1273	0.1370	0.1178	0.2483	0.2867	0.1507	0.1455
C. florida-1	0.3018	0.3417	0.3435	0.3049		0.0079	0.0079	0.1718	0.2051	0.0742	0.0697
C. florida-2	0.2987	0.3381	0.3399	0.3018	0.0504		0.0159	0.1723	0.2057	0.0831	0.0785
C. capitata	0.2682	0.3065	0.3081	0.2709	0.0252	0.0250		0.1615	0.1942	0.0654	0.0610
C. oblonga	0.7020	0.6459	0.7445	0.7445	0.5260	0.5199	0.4839		0.0240	0.1816	0.1962
C. alternifolia	0.6780	0.6231	0.7191	0.7191	0.5646	0.5578	0.5204	0.0768		0.2153	0.2305
C. chinensis	0.2662	0.3231	0.3248	0.2870	0.2629	0.1942	0.2309	0.8270	0.6791		0.0158
C. eydeana	0.1840	0.2349	0.2361	0.2022	0.1819	0.1200	0.1522	0.6712	0.6578	0.0515	

Ka: above diagonal; Ks: below diagonal; *C. suecica*-1 from Alaska, USA; *C. suecica*-2 from Norway. *C. florida*-1 as voucher of 02-16 (Xiang 2002) from Veracruz, Mexico; *C. florida*-2 as voucher of 02-36 (Fan 2002) from North Carolina, USA.

Shade: Ks values below 0.05 are not used to calculate the ratio of Ka/Ks.

Table 11. Pairwise nonsynonymous (Ka) and synonymous (Ks) substitution rates (the number of Ka and Ks per site) between 11 *Cornus* samples from the C-terminal domain. Estimated using Nei-Gojobori method with Jukes-Cantor model; number of sites = 105.

	C. canadensis	*C. unalaschkensis*	*C. suecica-1*	*C. suecica-2*	*C. florida-1*	*C. florida-2*	*C. capitata*	*C. oblonga*	*C. alternifolia*	*C. chinensis*	*C. eydeana*
C. canadensis		0.0040	0.0081	0.0081	0.0718	0.0763	0.0629	0.1157	0.1525	0.0121	0.0586
C. unalaschkensis	0.0000		0.0040	0.0040	0.0674	0.0719	0.0586	0.1111	0.1477	0.0162	0.0544
C. suecica-1	0.0152	0.0152		0.0000	0.0718	0.0763	0.0629	0.1157	0.1525	0.0203	0.0587
C. suecica-2	0.0152	0.0152	0.0000		0.0718	0.0763	0.0629	0.1157	0.1525	0.0203	0.0587
C. florida-1	0.1676	0.1669	0.1863	0.1863		0.0040	0.0081	0.1087	0.1452	0.0672	0.0457
C. florida-2	0.1667	0.1660	0.1853	0.1853	0.0000		0.0122	0.1136	0.1454	0.0717	0.0501
C. capitata	0.1499	0.1493	0.1683	0.1683	0.0307	0.0305		0.0993	0.1352	0.0584	0.0372
C. oblonga	0.3167	0.3152	0.2940	0.2940	0.2207	0.2194	0.1832		0.0377	0.1131	0.0858
C. alternifolia	0.3230	0.3216	0.3005	0.3005	0.2274	0.2261	0.1901	0.0594		0.1447	0.1211
C. chinensis	0.0801	0.0798	0.0970	0.0970	0.1690	0.1681	0.1328	0.3083	0.2927		0.0542
C. eydeana	0.1858	0.1850	0.2049	0.2049	0.1124	0.1118	0.0789	0.1815	0.1882	0.1315	

Ka: above diagonal; Ks: below diagonal; *C. suecica*-1 from Alaska, USA; *C. suecica*-2 from Norway. *C. florida*-1 as voucher of 02-16 (Xiang, 2002) from Veracruz, Mexico; *C. florida*-2 as voucher of 02-36 (Fan, 2002) from North Carolina, USA.

Shade: Ks values below 0.05 are not used to calculate the ratio of Ka/Ks.

Table 12. Ratio of Ka/Ks between 11 *Cornus* samples for the entire *myc*-like anthocyanin regulatory gene.

	C. canadensis	*C. unalaschkensis*	*C. suecica-1*	*C. suecica-2*	*C. florida-1*	*C. florida-2*	*C. capitata*	*C. oblonga*	*C. alternifolia*	*C. chinensis*	*C. eydeana*
C. canadensis											
C. unalaschkensis	NA										
C. suecica-1	NA	NA									
C. suecica-2	NA	NA	NA								
C. florida-1	0.399	0.394	0.430	0.433							
C. florida-2	0.401	0.412	0.437	0.450	NA						
C. capitata	0.423	0.435	0.484	0.497	0.413	0.378					
C. oblonga	0.401	0.384	0.406	0.408	0.380	0.379	0.393				
C. alternifolia	0.386	0.398	0.417	0.427	0.395	0.386	0.393	0.369			
C. chinensis	0.359	0.365	0.400	0.410	0.316	0.311	0.355	0.401	0.438		
C. eydeana	0.415	0.413	0.448	0.457	0.360	0.355	0.407	0.444	0.464	0.517	

C. suecica-1 from Alaska, USA; *C. suecica*-2 from Norway. *C. florida*-1 as voucher of 02-16 (Xiang 2002) from Veracruz, Mexico; *C. florida*-2 as voucher of 02-36 (Fan 2002) from North Carolina, USA.

NA: not applicable due to the value of Ks less than 0.05.

Table 13. Ratio of Ka/Ks between 11 *Cornus* samples for the interaction domain and the bHLH domain of *myc*-like anthocyanin regulatory gene.

	C. canadensis	*C. unalaschkensis*	*C. suecica-1*	*C. suecica-2*	*C. florida-1*	*C. florida-2*	*C. capitata*	*C. oblonga*	*C. alternifolia*	*C. chinensis*	*C. eydeana*
C. canadensis		NA	NA	NA	0.255	0.235	0.240	0.123	0.105	0.224	0.235
C. unalaschkensis	NA		NA	NA	0.217	0.319	0.318	0.157	0.136	0.268	0.284
C. suecica-1	0.499	0.291		NA	0.253	0.347	0.344	0.173	0.147	0.287	0.305
C. suecica-2	0.763	0.396	NA		0.253	0.397	0.386	0.199	0.181	0.314	0.334
C. florida-1	0.274	0.267	0.371	0.405		NA	NA	0.119	0.105	0.204	0.212
C. florida-2	0.307	0.297	0.403	0.454	0.157		NA	0.041	0.026	0.101	0.098
C. capitata	0.275	0.269	0.382	0.435	NA	NA		0.086	0.064	0.161	0.165
C. oblonga	0.287	0.327	0.334	0.334	0.327	0.331	0.334		0.044	0.176	0.156
C. alternifolia	0.346	0.398	0.399	0.399	0.363	0.369	0.373	0.313		0.147	0.114
C. chinensis	0.411	0.352	0.464	0.525	0.282	0.428	0.283	0.220	0.317		NA
C. eydeana	0.568	0.463	0.616	0.720	0.383	0.654	0.401	0.292	0.350	0.307	

Ka/Ks for interaction domain: above diagonal; Ka/Ks for bHLH domain: below diagonal. *C. suecica*-1 from Alaska, USA; *C. suecica*-2 from Norway. *C. florida*-1 as voucher of 02-16 (Xiang 2002) from Veracruz, Mexico; *C. florida*-2 as voucher of 02-36 (Fan 2002) from North Carolina, USA.

NA: not applicable due to the value of Ks less than 0.05.

Table 14. Ratio of Ka/Ks between 11 *Cornus* samples for the Acidic domain and the C-terminal domain of *myc*-like anthocyanin regulatory gene.

	C. canadensis	*C. unalaschkensis*	*C. suecica-1*	*C. suecica-2*	*C. florida-1*	*C. florida-2*	*C. capitata*	*C. oblonga*	*C. alternifolia*	*C. chinensis*	*C. eydeana*
C. canadensis		NA	NA	NA	0.399	0.401	0.517	0.511	0.498	0.433	0.464
C. unalaschkensis	NA		NA	NA	0.402	0.404	0.519	0.505	0.492	0.421	0.452
C. suecica-1	NA	NA		NA	0.475	0.454	0.655	0.557	0.537	0.484	0.516
C. suecica-2	NA	NA	NA		0.453	0.437	0.619	0.539	0.525	0.463	0.495
C. florida-1	0.428	0.404	0.385	0.385		NA	0.388	0.493	0.481	0.291	0.342
C. florida-2	0.458	0.433	0.412	0.412	NA		0.398	0.492	0.473	0.298	0.346
C. capitata	0.420	0.392	0.374	0.374	NA	NA		0.523	0.505	0.424	0.469
C. oblonga	0.365	0.352	0.394	0.394	0.493	0.518	0.542		NA	0.559	0.573
C. alternifolia	0.472	0.459	0.507	0.507	0.638	0.643	0.711	0.635		0.561	0.589
C. chinensis	0.151	0.203	0.209	0.209	0.398	0.427	0.440	0.367	0.494		NA
C. eydeana	0.315	0.294	0.286	0.286	0.407	0.448	0.471	0.473	0.643	0.412	

Ka/Ks for acidic domain: above diagonal; Ka/Ks for C-terminal domain: below diagonal. *C. suecica*-1 from Alaska, USA; *C. suecica*-2 from Norway. *C. florida*-1 as voucher of 02-16 (Xiang 2002) from Veracruz, Mexico; *C. florida*-2 as voucher of 02-36 (Fan 2002) from North Carolina, USA.

NA: not applicable due to the value of Ks less than 0.05.

Table 15. 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 (Ka/Ks)	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; Ka: nonsynonymous substitution rate; Ks: nonsynonymous substitution rate. See M & M and Yang et al. (2000) for the definitions of parameters

Table 16. Likelihood ratio test comparing models of variable ω ratios among sites

Comparisons	Log-likelihood values	Degree of Freedom (υ)	χ^2 distribution at υ	LRT statistic
M1 vs M2	-5965.77 vs -5962.58	2	*P*0.05=5.99; *P*0.01=9.21	6.38*
M0 vs M3	-5980.75 vs -5962.56	4	*P*0.01=13.28; *P*0.005=14.86	36.38**
M7 vs M8	-5962.97 vs -5962.56	2	*P*0.1=4.61; *P*0.05=5.99	0.82

* Significant difference at *P*<0.05 level; **Significant difference at *P*0.01 level; *** Significant difference at *P*>0.005 level.

Table 17. The number and distribution of positive selection sites (with ω>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 18. Pattern of positive selection sites with *P*>99% (ω>1) suggested by Codon-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

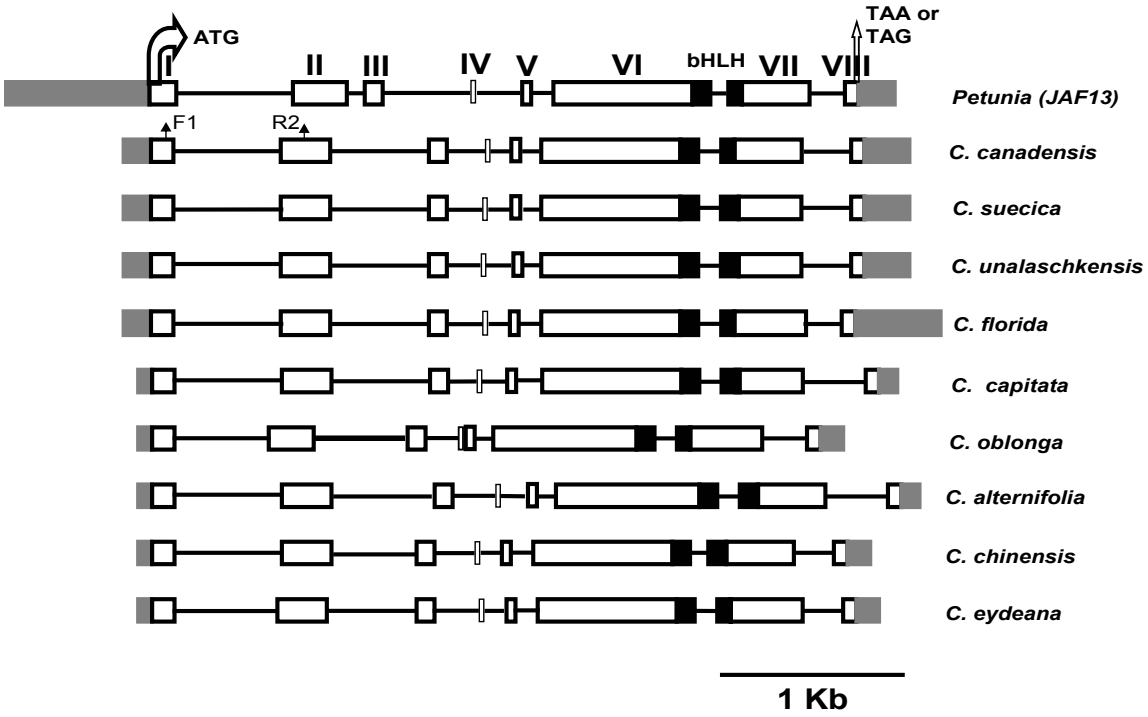


Fig. 1. Schematic map showing the overall structure of the *myc*-like anthocyanin regulatory gene (*R* homologue) for *Cornus* and *Petunia hybrida* (*JAF13*) as deduced from the full length of nucleotide sequences. The position of the ATG translation start and the TAA or TAG translation stop codons are indicated. The box represents the exon, and line stands for intron. Eight exons are ordered using Roma numerals (I-VIII). Seven insertion-deletions among *Cornus* species were found at exon VI through exon VIII (see Table 2 for details). The region encoding bHLH is shown in dark which locates 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 number $$$) 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).

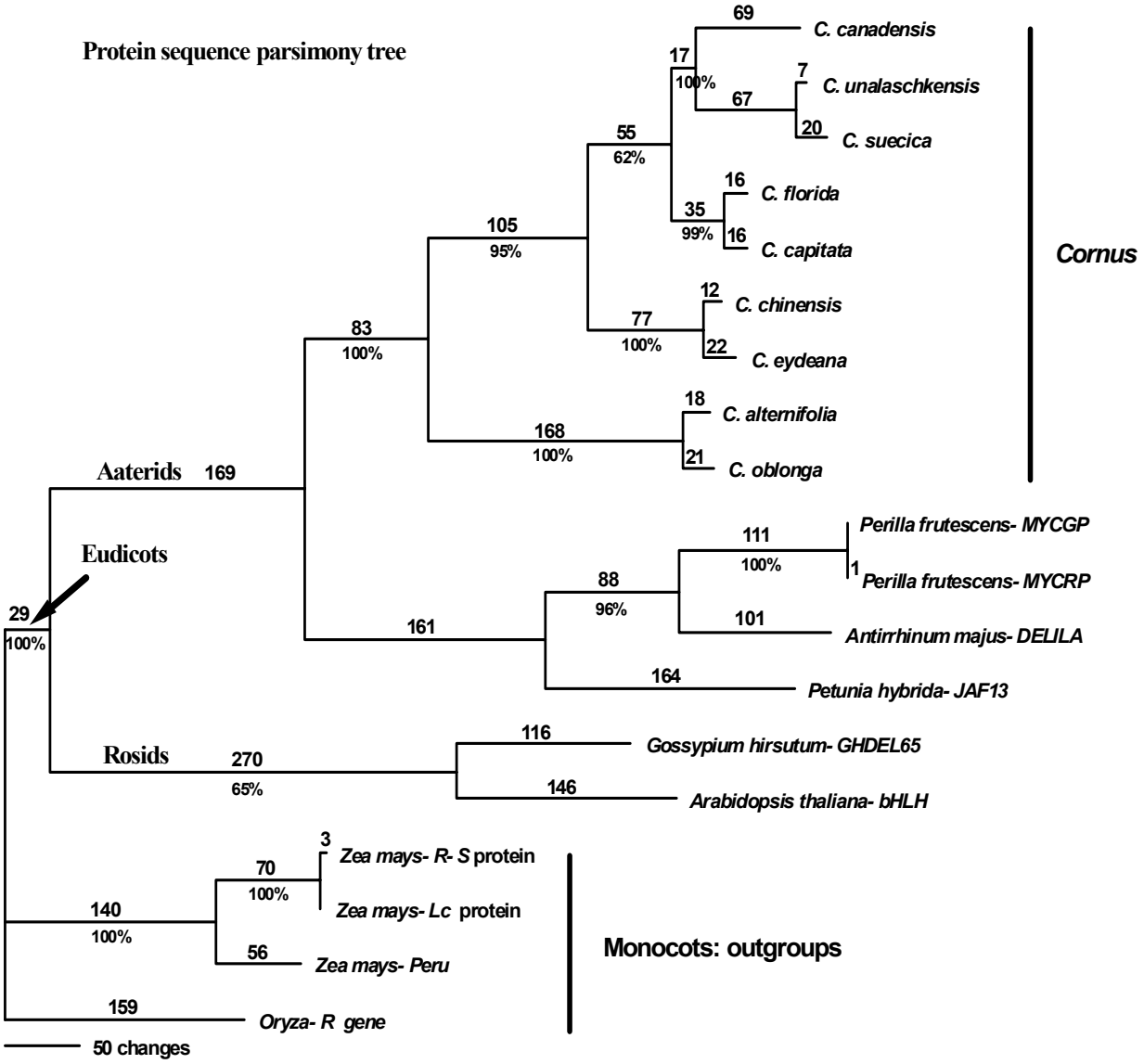


Fig. 2. One of four parsimonious trees of amino acid sequences of the *myc*-like anthocyanin regulatory gene (*R*) from *Arabidopsis*, *Petunia*, *Perilla*, *Gossypium*, *Zea mays*, *Oryza*, *Antirrhinum majus*, and *Cornus*. The sequences of *Cornus* were generated in this study, and all other sequences 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 number AF013465); *Gossypium hirsutum ghdel* (Matz and Burr 2001, unpublished. Genbank accession number AF336280); *Petunia hybrida-jaf*13

(Quattrocchio et al. 1998. Genbank accession number AF020545). *Antirrhinum majus-delila* (Goodrich, Carpenter, and Coen 1992. Genbank accession number M84913). *Perilla frutescens-mycrp*, *mycgp* (Gong et al. 1999. Genbank accession number AB024050-Myc-rp, AB024051-Myc-gp). *Zea mays* (*Lc*, Ludwig et al. 1989. Genbank accession number M26227; Zm *B-Peru*, Radicella et al. 1991, Genbank accession number X57276; *R-S*-protein, Perrot and Cone 1989, Genbank accession number X15806); *Oryza-R* (Hu et al. 1996. Genbank accession number U39860). The sequences from maize and rice were treated as outgroups. The analysis was performed on PAUP4.0b10. Base substitutions are indicated above branches; bootstrap values are marked below branches. Tree length = 2060; CI = 0.839; RI = 0.828.

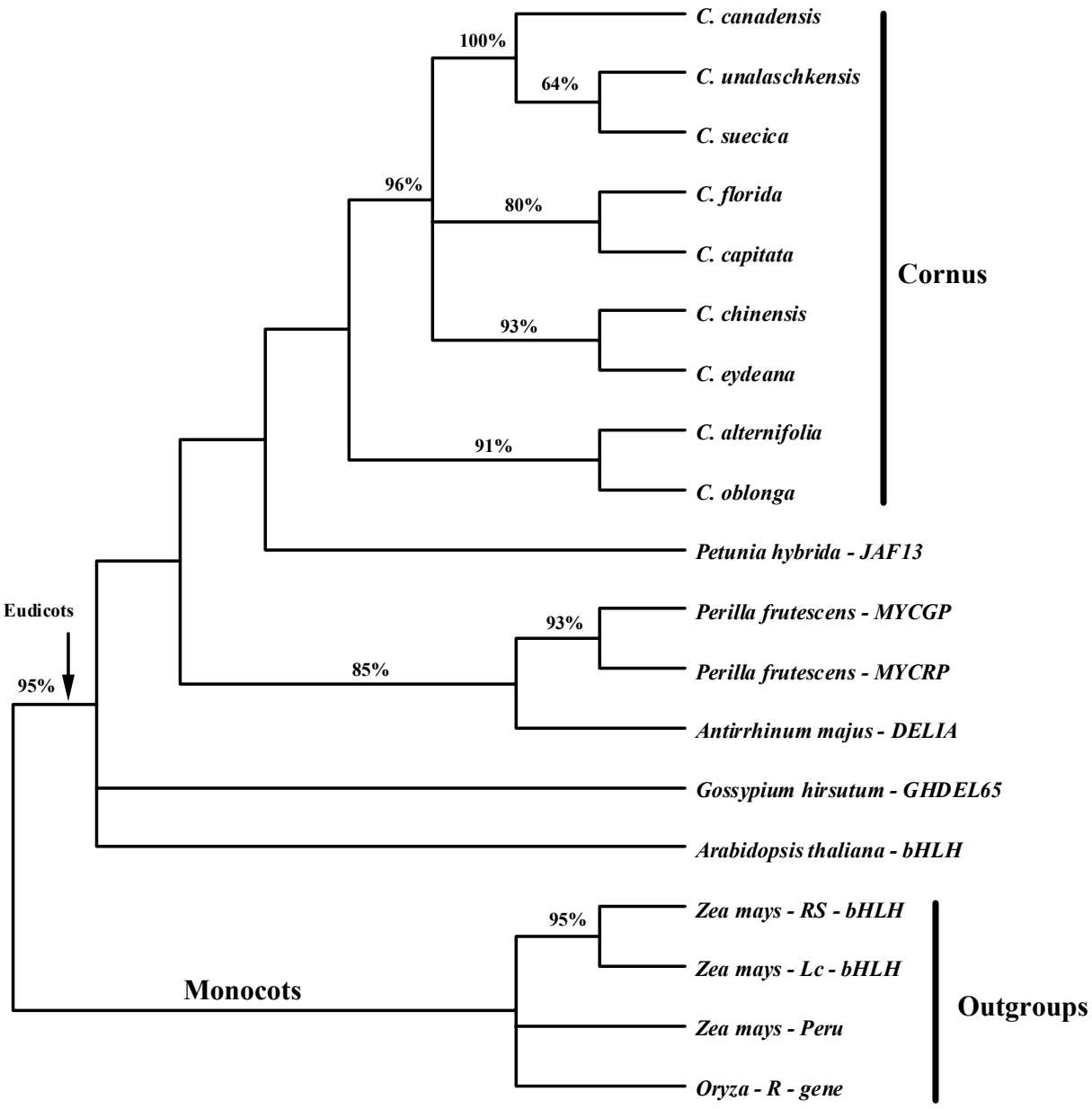


Fig. 3. Strict consensus of twelve parsimonious trees of just bHLH domain of the *myc*-like anthocyanin regulatory gene from *Arabidopsis*, *Petunia*, *Perilla*, *Gossypium*, *Zea mays*, *Oryza*, *Antirrhinum majus,* and *Cornus* (see Fig. 2 for detailed information for genes). The analysis was performed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

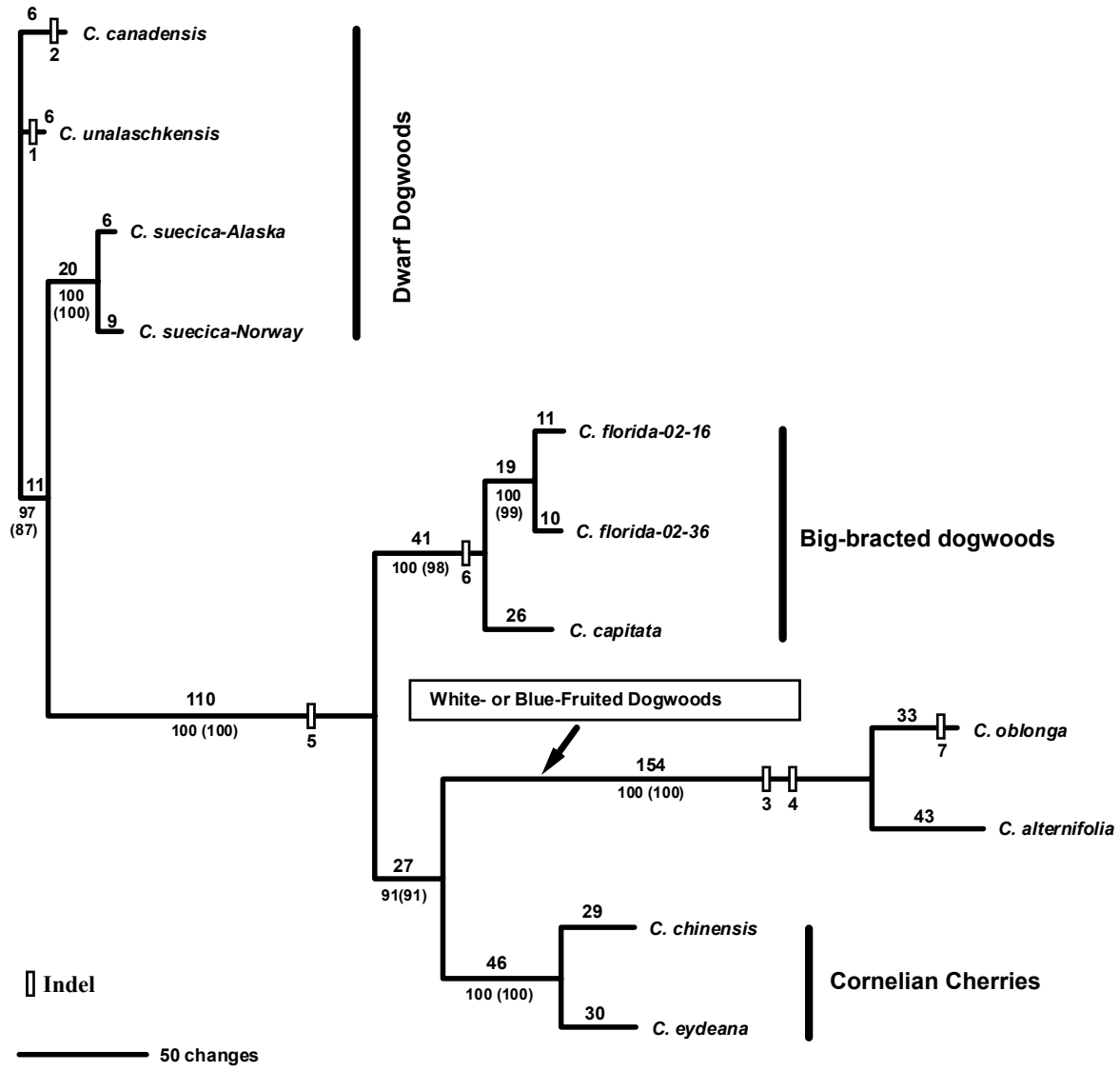


Fig. 4. The single parsimonious unrooted tree inferred from exon sequences of eleven taxa of nine *Cornus* species (tree length = 632; CI = 0.877; RI = 0.899). Base substitutions are indicated above branches; bootstrap values for both parsimony and ML tree (parentheses) are marked below branches. Seven indels identified from exons are marked (see Table 2 for the details).

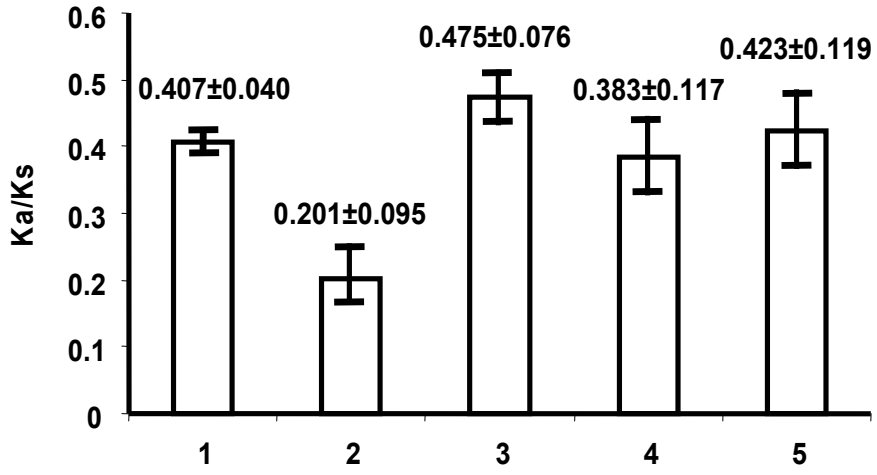
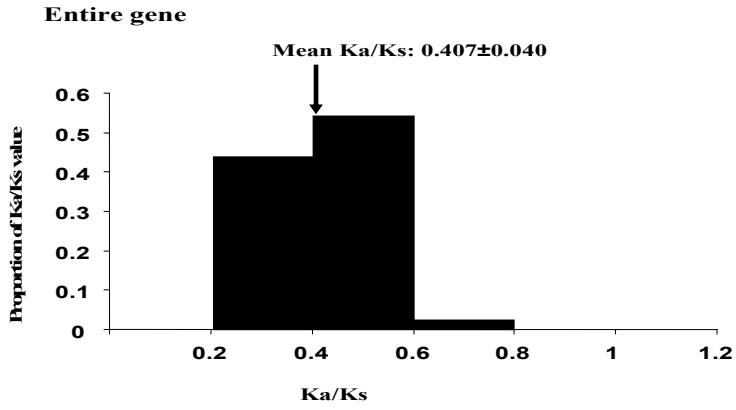
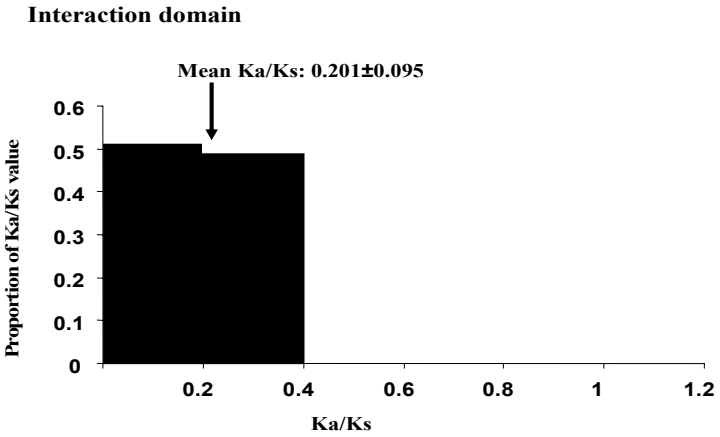
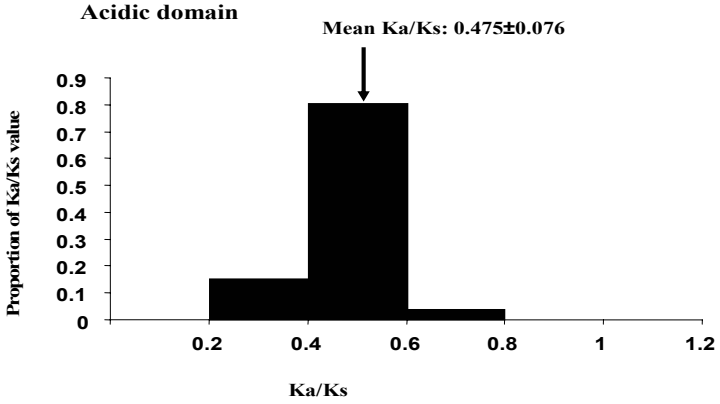


Fig. 5. Mean and standard deviation of ratio of Ka/Ks between nine species of *Cornus*. 1. Entire gene; 2. Interaction domain; 3. Acidic domain; 4. bHLH domain; 5. C-terminal domain.


Entire gene
Mean Ka/Ks: 0.407±0.040
Proportion of Ka/Ks value
0.6
0.5
0.4
0.3
0.2
0.1
0
0.2
0.4
0.6
0.8
1
1.2
Ka/Ks


Interaction domain
Mean Ka/Ks: 0.201±0.095
Proportion of Ka/Ks value
0.6
0.5
0.4
0.3
0.2
0.1
0
0.2
0.4
0.6
0.8
1
1.2
Ka/Ks


Acidic domain
Mean Ka/Ks: 0.475±0.076
Proportion of Ka/Ks value
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.2
0.4
0.6
0.8
1
1.2
Ka/Ks

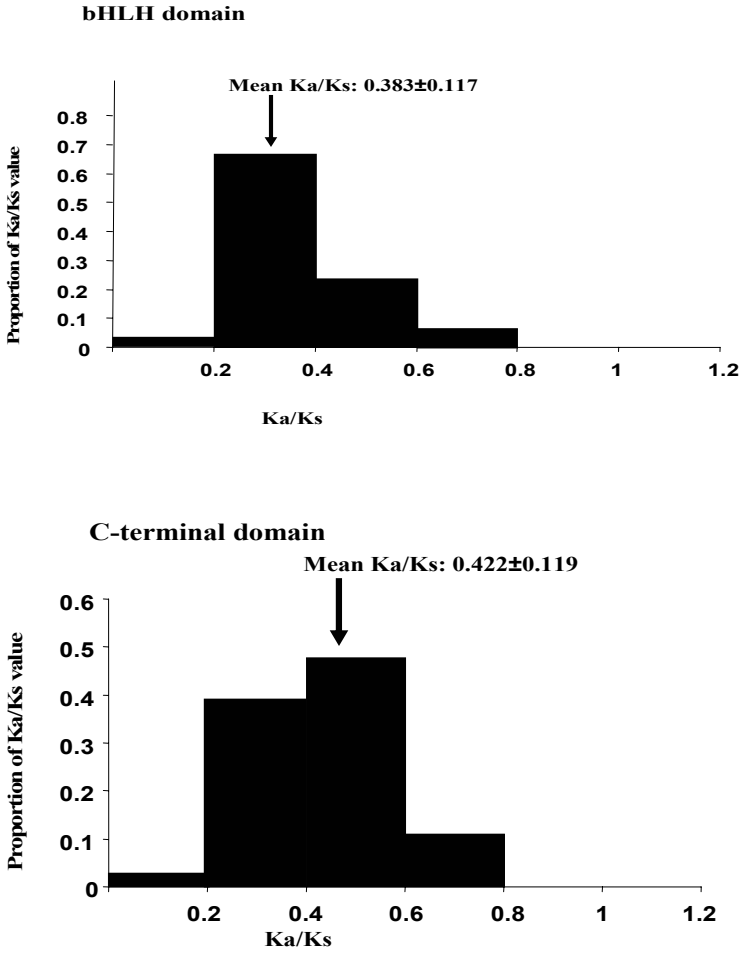
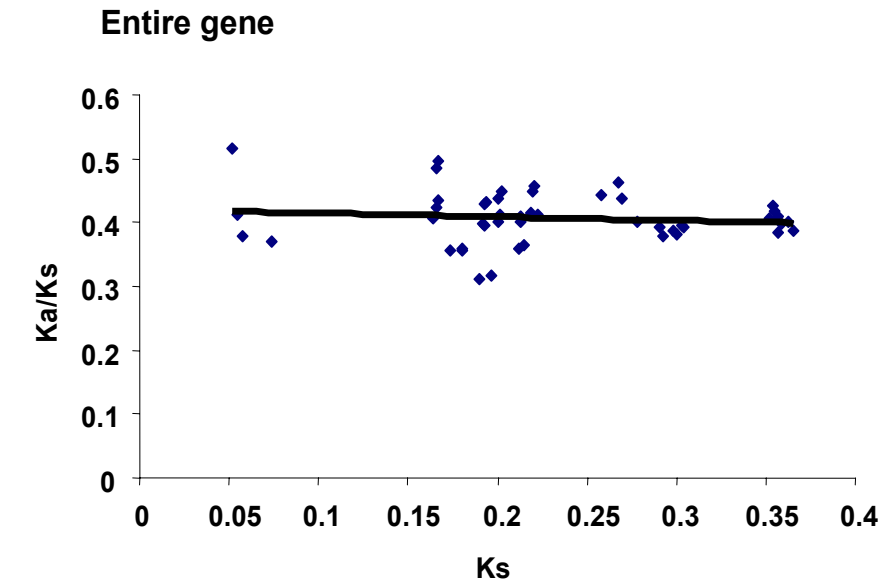
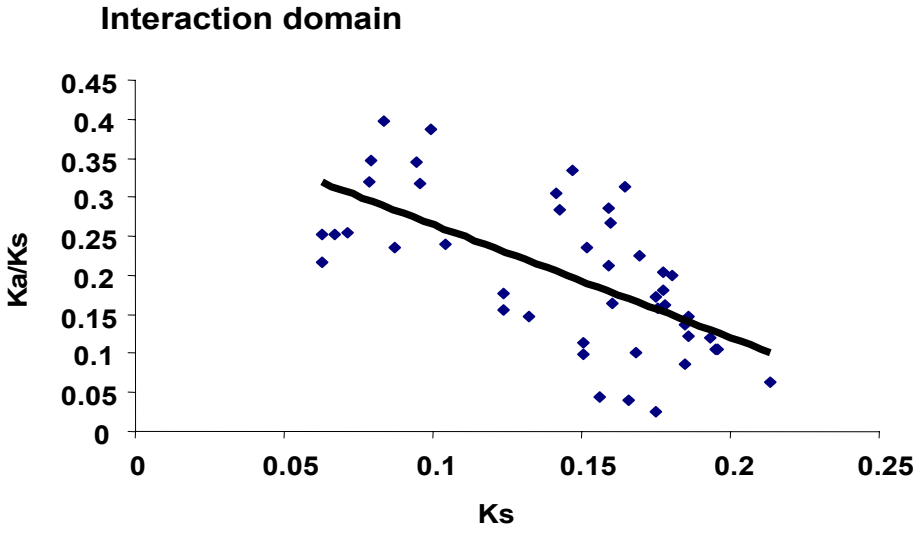
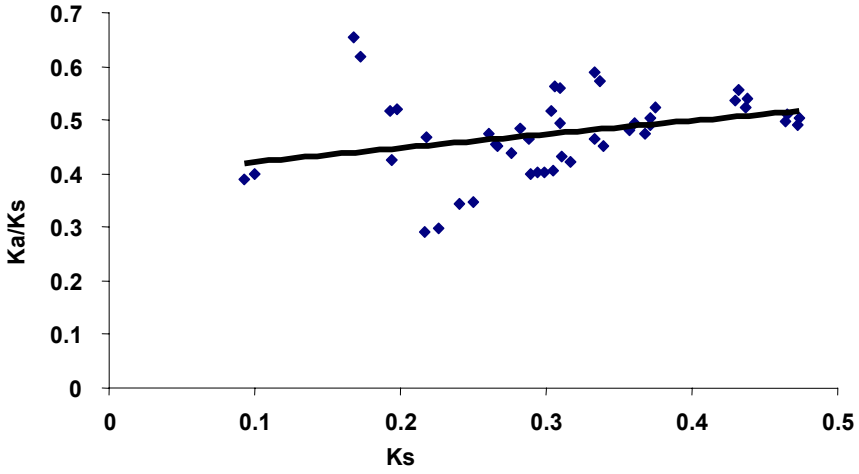


Fig. 6. Distribution of ratio of Ka/Ks for entire gene, interaction domain, acidic domain, bHLH domain, and C-terminal domain for 11 samples of nine *Cornus* species. Mean values and standard deviations are indicated. Pairwise comparisons that had synonymous substitutions below 0.05 are not shown in the histograms and were not included in the analyses.

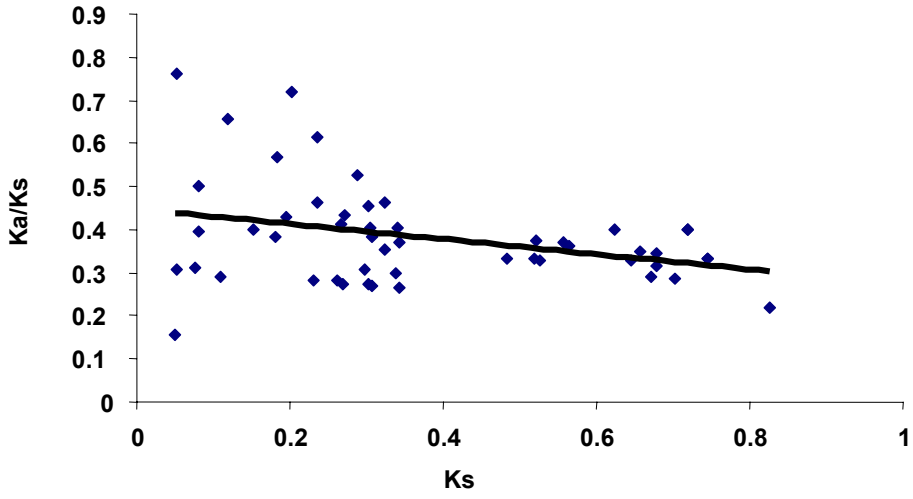

Entire gene
Ka/Ks
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
Ks


Interaction domain
Ka/Ks
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.05
0.1
0.15
0.2
0.25
Ks

Acidic domain



bHLH domain



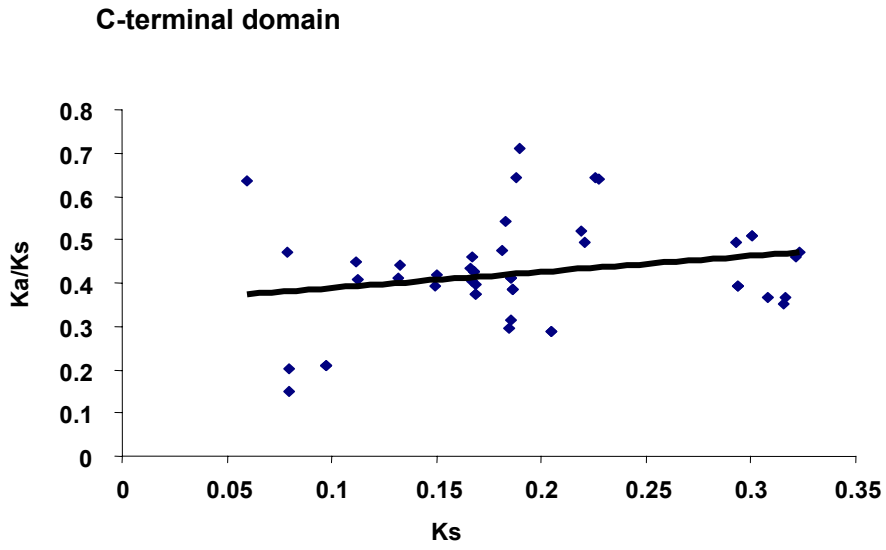


Fig. 7. Plots of Ka/Ks versus Ks for the entire gene, interaction domain, acidic domain, bHLH domain, and C-terminal domain for 11 *Cornus* species.

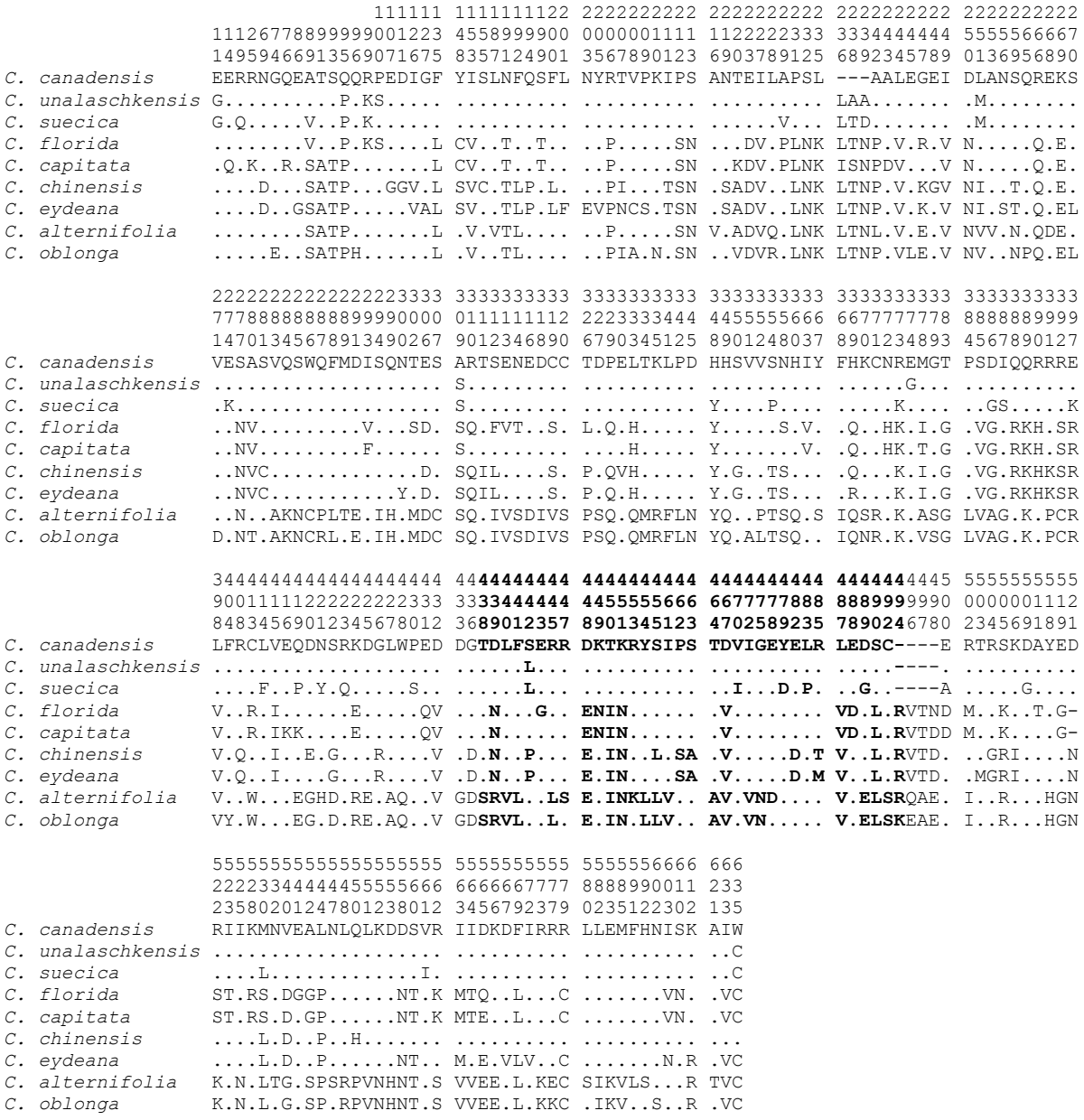


Fig. 8. Aligned variable amino acid sites of the *myc*-like anthocyanin regulatory gene among species of *Cornus*. Dots indicate identity to the topmost sequence. Dashes represent gaps. Interaction domain: Sites 1-194; acidic domain: sites 195-436; bHLH domain: 437-494; C-terminal domain: 532-637. bHLH domain is bolded.

Chapter IV

Linking the *myc*-Like Anthocyanin Regulatory Gene Sequence Variation to Phenotypic Diversity in Petal Colorization in the Dwarf Dogwood Species Complex (*Cornus*): Accelerated Gene Evolution and Positive Selection

Authors: Chuanzhu Fan*, Michael D. Purugganan[†], Xi Chen[‡], Brian M. Wiegmann[§], and (Jenny) Qiu-Yun Xiang*

*Department of Botany, North Carolina State University, Raleigh, NC 27695-7612

[†]Department of Genetics, North Carolina State University, Raleigh, NC 27695-7614

[‡]Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203

[§]Department of Entomology, North Carolina State University, Raleigh, NC 27695-7613

Date of receipt:

Key words: Anthocyanin regulatory gene; *Cornus*; Hybridization; Gene evolution; Nucleotide polymorphism

Corresponding author: Chuanzhu Fan, Department of Botany, Campus box 7612, North Carolina State University, Raleigh, NC 27695-7612; fax number: 919-515-3436; email: cfan3@unity.ncsu.edu

Manuscript in preparation: Follow the format of '*Molecular Ecology*'

Chuanzhu did all of the work reported in this chapter. Xi Chen assisted with GLM test. Drs. Michael D. Purugganan, Brian M. Wiegmann, and (Jenny) Qiu-Yun Xiang provided scientific advice and guidance.

Abstract

Hybridization and polyploidization are important evolutionary forces in plant speciation. Hybrids, and especially allo-polyploids, derived from hybridization of divergent species, have long been recognized to be linked with novel morphological and ecological adaptations, but how genomic evolution ultimately translates into novel evolutionary opportunity remains unknown. Some recent studies suggest that regulatory genes in hybrids and polyploids play a key role in the processes generating novel genotypes and morphological or physiological phenotypes. Here, we examine the evolution of the *myc*-like anthocyanin regulatory gene in the dwarf dogwoods species complex to explore the pattern of gene evolution and its correlation to phenotypic variation in petal colors. Full length nucleotide sequences of the *myc*-like anthocyanin regulatory gene were sequenced and characterized for 47 dwarf dogwood samples representing 44 geographical sites. We performed sequence diversity and neutrality tests to understand the pattern of gene evolution and the evolution processes. We also tested the association between amino acid sequence and variation of petal color using a general linear model (GLM) with multiple regressions. Genealogy of haplotypes was constructed to elucidate the evolutionary history of the gene in the complex. Our results suggest that excess of low-frequency polymorphisms exist in the species complex, implying accelerated gene evolution. Positive selection and rapid population expansion are likely to be the main forces acting on the elevated gene evolution. Our results also reveal that there is significant correlation between variable sites in the protein, particularly in the acidic and bHLH domains, and the variation of petal colors. Both nucleotide diversity tests and gene genealogy suggest that gene flow, recombination, and

introgression occur within dwarf dogwoods, indicating a dynamic evolutionary process after hybridization.

Introduction

Hybridization and polyploidization are important evolutionary processes in plants and have long been considered significant forces in plant evolution (Rieseberg & Wendel 1993; Arnold, 1997; Wendel 2000). It is estimated that the origins of 50-70% of angiosperms involved past natural hybridization and polyploidization (Stebbins 1971; Grant 1981; Masterson 1994; Soltis & Soltis 1999; Soltis & Gitzendanner 1999). Recent interdisciplinary approaches and studies combining morphological, phylogenetic, and molecular genetic perspectives have greatly enhanced our understanding of the genetic and evolutionary consequences of hybridization and polyploidization (see Rieseberg 1997; Arnold 1997; Soltis & Soltis 1999; Otto & Whitton 2000; Wendel 2000). However, our knowledge regarding the physio-ecological consequences of genomic changes at the molecular level following hybridization and polyploidization is still very limited (but see Brochmann et al. 1992; Wendel et al. 1995; Song et al. 1995; Soltis et al. 1995; Cronn et al. 1999). Studies of regulatory gene evolution in polyploids and hybrids are critical to the development of an increased understanding of the ecological consequences of genomic evolution in these plants. Such studies will also provide clues to mechanisms of molecular-driven morphological diversification in hybrids and polyploids.

Hybrids, and especially allo-polyploids, derived from hybridization of divergent species, are often responsible for the origins of novel morphological and ecological

adaptations (Wendel 2000), but how gene evolution ultimately translates into novel evolutionary opportunity is an open area for research and still only poorly known. Some recent studies suggest that regulatory genes in hybrids and polyploids play a key role in the processes generating novel genotypes and morphological or physiological phenotypes upon which natural selection might act (reviewed by Wendel 2000). Regulatory genes represent a class of loci that control the expression of other genes. It is well known that regulatory gene evolution can be a significant factor in organismal diversification (Dickinson 1988; Doebley 1993). Changes in regulatory loci, for example, are thought to underlie the evolution of developmental mechanisms that result in morphological differentiation between taxa (Carroll 1995; Palopoli & Patel 1996; Purugganan 1998, 2000). Some recent studies also suggest that regulatory genes are better candidates for tracking speciation events than structural genes (e.g. Barrier et al. 1999, 2001). They are, therefore, valuable markers for studying evolutionary history and speciation of organisms.

Various studies indicated that the expression of anthocyanin regulatory genes and their homologues in *Zea mays*, *Nicotiana*, *Arabidopsis*, *Petunia*, *Antirrhinum majus, Gossypium*, and *Solanum lycopersicum* activate the structural genes and induce pigmentation in a wide variety of tissues, especially in flowers, and that mutations of the *myc*-like regulatory genes result in partial expression of anthocyanin pigments in petals (Ludwig & Wessler 1990; Radicella et al. 1991; Martin et al. 1991; Goodrich et al. 1992; Consonni et al. 1992, 1993; Lioyd et al. 1992; Quattrocchio et al. 1993, 1998; Goldsbourough et al. 1996; Hu et al. 1996; De-Vetten et al. 1997). The study of *myc*-like anthocyanin regulatory genes in hybrids and polyploids with different petal color phenotypes may thus be particularly useful for

describing the connection between molecular evolution and phenotypic differentiation after hybridization and polyploidization.

Dwarf dogwoods or bunchberries (*Cornus* Subg. *Arctocrania*) are perennial rhizomatous ground cover herbs. Three species, C. *canadensis* L. (defined as Lineage 'A'), *C. suecica* Lamark (1786) (defined as Lineage 'E'), and *C. unalaschkensis* Ledebour (1844) (defined as Lineage 'C') have been described. The first two species are diploid (2n=22) and the third is a tetraploid (2n=44) (Bain & Denford 1979). The species complex is distributed circumboreally in the Northern Hemisphere and exhibits considerable intermediacy between morphological extremes displayed in *C. canadensis* and *C. suecica*, the two diploid species. *Cornus canadensis* has white/cream petals. *Cornus suecica*, in contrast, has dark purple petals. These two diploid species also occur in different habitats. In the Pacific Northwest, *C. suecica* occurs in completely open areas at high elevations (from the transitional zone to above the tree line) where wind is strong and UV light is intense. *C. canadensis* occurs at lower elevations at the edge or inside the coniferous forests in sunny and relatively open areas. Intermediate forms include the allotetraploid, *C. unalaschkensis*, presumably derived from *C. suecica* and *C. canadensis*, and two other informal lineages, *C. canadensis* > *C. suecica* (defined as Lineage 'B') and *C. canadensis* < *C. suecica* (defined as Lineage 'D') (Murrell 1994). These plants have different combinations of morphological characters of the two diploid species and their petals are bicolor, with the apical part being purple and the basal portion being white. The purple portion of petals varies from 1/8 to nearly completely purple. These intermediate forms occur across a range in elevation inhabited by the two diploid species. Studies from cytology and morphology suggest that morphological intermediacy in this group is the result of extensive hybridization and introgression between

the two diploid species and may also involve the tetraploid species (Bain & Denford 1979; Murrell 1994). Therefore, this species complex provides a good system in which to study the evolution of anthocyanin regulatory genes and their potential phenotypic consequences. In the present study, the full length nucleotide sequence (~4000 bp) of the *myc*-like anthocyanin regulatory gene was sequenced for 47 samples representing 44 different geographical areas (populations). The goals of this study are: (1) to examine the rate and pattern of gene variation in the complex; and (2) to determine the relationship between sequence variation and petal color.

Materials and Methods

Sampling

Samples of dwarf dogwoods were collected from the Pacific Northwest region of North America where *C. unalaschkensis* and the other four taxa co-occur. Our strategy was to sample extensively in this region to include multiple populations of each taxon.

Plant samples were collected from 43 sites in this area, and were placed into five lineages based on morphology. The sampling sites span geographical areas of Idaho (ID, USA), Oregon (OR, USA), Washington (WA, USA), Alaska (AK, USA), British Columbia (BC, Canada), and the Yukon (YK, Canada) (Fig. 1). Multiple plants at each locality were collected. Because these plants are rhizomatous perennial herbs, it is unclear which adjacent samples represent truly different individuals or might just be different "shoots" of the same plant. Therefore, only one plant from each locality was used in this study. One additional

sample of *C. suecica* from Norway was also included to increase the sampling diversity (Table 1).

DNA extraction and morphological identification

Total genomic DNA was extracted from fresh leaves, using a previously described protocol (see Xiang et al. 1998). Petal color was recorded into five categories: purple, 2/3 purple, 1/2 purple, 1/3 purple, and white. Among 43 populations, three have completely or mostly purple petal color (# 27, 29, and 43). The plants with completely white petals include fourteen populations (# 6, 7, 8, 14, 15, 16, 17, 18, 19, 20, 21, 32, 33, and 40). The others have bicolor petals, with the ratio of purple/white in their petal ranging from 1/3 to 2/3 (Table 1). All samples were identified into one of five lineages based on several morphological diagnostic features described in a previous study (Murrell 1994; Tables 1).

PCR Amplification, Thermal asymmetric interlaced (TAIL) PCR, sequencing and cloning

In order to amplify DNA sequences of this gene in *Cornus*, degenerate primers (F1 and R2) were designed from published sequences of *myc*-like anthocyanin regulatory gene sequences of dicots including *Arabidopsis*, *Petunia*, *Gossypium hirsutum*, *Perilla*, *Antirrhinum majus* (forward primer F1-CAATGGAGYTATRTYTTHTGGTC and reverse primer R2-TCRGTRAGRTCTTCWGGDGATAATGC). 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). These two primers were used for initial PCR and sequencing for ten samples from five lineages. The PCR reaction using these two primers generated a 800-bp length fragment. Cloning and Blast analysis of this PCR product revealed two types

(designated as Type 'A' and Type 'B') of sequences, both of which are highly similar to the anthocyanin regulatory gene in *Petunia*, *Perilla*, and *Antirrhinum majus*. Type-specific primers were subsequently designed to amplify and sequence a single type (type 'A') (Table 2). Type 'A' sequences were elongated via sequential PCR reactions using sequential locus specific forward primers from known sequences and locus specific/degenerate reverse primers from developed alignment of Genbank sequences. Sequences of flanking regions and ends of this gene were achieved using thermal asymmetric interlaced (TAIL) PCR (Liu and Whittier 1995; Liu et al. 1995; Table 3). The entire nucleotide sequence of this gene was obtained for ten samples representing five lineages of dwarf dogwoods using methods described above. Dogwood "universal" locus specific primers were then designed based on these sequences.

Using the "universal primers" (Table 2), the entire nucleotide sequence of the *myc*-like anthocyanin regulatory gene was amplified for the remaining thirty-seven samples. The primer combinations used are: F0A-R2A2, F2A (or F2A1)-R3', F4A-R4A, F6A-R7A, and F7A2-R9A (Table 2). For all these primer combinations, the adjacent fragments amplified overlap by at least 50 base pairs at two ends. Gel extraction procedure (1.5% agarose gel electrophoresis and then purified using QIAquick PCR purification kit from Qiagen, Maryland 20874, USA) or TOPO TA cloning was applied to some cases where multiple PCR bands were obtained. In all cases, both DNA strands were sequenced. All degenerate primers and locus specific primers described above were 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).

PCR reactions contained the following: 5 μL of 10x Mg^{2+} free buffer, 6 μL of 25mmol/L $MgCl_2$, 6-10 μL of 2.5mmol/L dNTPs, 0.5μL of 20μmol/L forward primer, 0.5μL of 20μmol/L reverse primer, 5μL of DMSO (dimethyl sulfoxide), 1-5 μL of BSA (Bovine serum albumin, 10mg/ml), 0.3μL of *Taq* polymerase (Promega), 5-10 μL of 20ng/μL total DNA extract, and calibrated to final 50μL using deionized water. In order to avoid non-specific primer annealing and increase yield of PCR products, a hot-start (six minutes 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 seconds for one cycle; (2) 30-40 cycles of 94°C for 45 sec, 50-60°C (annealing temperature optimized based on the melting temperature (Tm) of primers, Table 2) for 1 min, 72°C for 1.5-2.5 min; (3) a terminal phase at 72°C for 6 min.

TOPO TA cloning was used to isolate the different types of sequences for the initial PCR amplification using primer F1 and R2. The PCR products were purified and cloned to competent *E. coli* cells using TOPO TA cloning techniques (Invitrogen Life Technologies, Carlsbad, California, USA 92008). The growing colonies were screened for positive transformants using PCR amplification by T3 and T7 primers located on 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, Wisconsin 53711-5399, USA). The purified plasmid DNA products were directly sequenced.

The double-stranded (DS) PCR products were cleaned using 20% PEG (polyethylene glycol) 8000/2.5 mol/L NaCl (Morgan & Soltis 1993; Soltis & 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, California, USA). Cycle-sequencing reactions (10 μL) were prepared by combining 2 μL terminator ready reaction mix, 2 μL sequencing buffer (200 mmol/L Tris-ph8.0, 5mmol/L $MgCl_2$), 0.6 μL primer (5 μmol/L), 2 μL or 4 μL of 200 ng/μL cleaned PCR products or plasmid DNA, 0.5 μL DMSO, and 2.9 μL (for PCR product reactions) or 0.9μL (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 96ºC for 30 sec, 50ºC for 15 sec, and 60ºC for 4 min. Products of cycle-sequencing were cleaned using ethanol/sodium acetate precipitation (ABI applied Biosystems, Foster City, California 94404, USA) with an additional 95% ethanol wash. The cleaned sequencing products were analyzed on an ABI-377 automated sequencer (Applied Biosystems, Foster City, California 94404, USA). The sequence chromatogram output files for all samples were checked and edited base by base manually before being aligned.

Test of sequence diversity and selection

To investigate sequence diversity within the dwarf dogwood species complex, comparisons were performed within and among three petal color phenotypic groups: group 'CC'- *C. canadensis* with white petals; group 'CS'- *C. suecica* with purple petal; group 'CH'- hybrids with bicolor petals. These groups were classified based on both petal colors and the sequence marker that distinguishes *C. suecica* and *C. canadensis* (see Table 4). Group 'CC' includes twelve samples (6-1, 7-1, 8-1, 14-1, 15-1, 16-1, 17-2, 18-3, 19-1, 21-2, 32-2, and 33-2) collected from various sites of the collecting area (see Fig. 1). Group 'CS' contains seven samples (27-1, 27-2, 29-1, 29-2, 43-1, 43-2, 94-388) from Alaska and

Norway. Group 'CH' includes 26 samples across a wide range of collecting area (see Fig. 1). Pattern and rates of sequence variation were assessed within and among the three groups. In addition, all samples were pooled together as a single group for a comparison. The sequence diversity test was performed using nucleotide diversity (π) (Nei 1987) and the population mutation parameter of Watterson's θ (denote here as θ_w) (Watterson 1975) in DNASP version 3.99 (Rozas et al. 2003). Tests of selection were conducted using Tajima's (1989) and Fu & Li's methods (1993) in DNASP. Both Tajima's and Fu & Li's tests compare two estimates of the mutation parameter $\theta = 4N\mu$, where N is the effective population size and μ is the mutation rate per sequence and per generation. The first value of θ is estimated using the average pairwise nucleotide difference (the value of π), and a second value of θ estimated using the number of segregating (or polymorphic) sites in the entire sample matrix of sequence diversity (the value of θ_w) (Tajima 1989). Both Tajima and Fu & Li tests of selection examine deviation from its neutral expectation by estimating the test statistic D for θ. Loci evolving neutrally have a value of zero for D. A positive D-value indicates a deficiency of low-frequency polymorphism, a consequence of purifying selection (deleterious mutation). Negative D values indicate an excess of low-frequency polymorphism, a result of diversifying positive selection, a recent selective sweep, or demographic factors such as population expansion (Lawton-Rauh et al. 2003). Significance of Tajima's D and Fu and Li's D were assessed using coalescent simulations of 2000 replicates incorporated with the number of segregating sites and the estimated recombination rate (γ). The value of γ was calculated using Hey & Wakeley's method implemented in the program of SITES (Hey & Wakeley 1997).

Association between Petal Color and Amino Acid Sequence Variation

The association between amino acid sequence variations and petal color patterns were tested using ANOVA (Analysis of Variance). A total of 38 informative variable amino acid sites were found among 47 samples (Fig. 3). Fifteen sites that have over 10% variation among 47 samples were used to test the association between variation of amino acid and diversity of petal color. Petal color was scored into five categories. 1 and 0 represent completely purple and whites petals, respectively. Bicolor petals were coded as 0.33 (1/3 purple + 2/3 white), 0.5 (1/2 purple and 1/2 white), and 0.67 (2/3 purple and 1/3 white). All data were analyzed using SAS® software (SAS Institute Inc. 1999). Three tests were performed. First, fifteen amino acid sites were considered as 15 random variables, and the General Linear Model (GLM) procedure was used to do multiple regressions to analyze the association between individual site and petal color. Secondly, using four pre-defined functional domains (Gong et al. 1999), all 15 amino acid sites were grouped by domains and compared using the same procedures as above. Finally, because the domain-groups analysis showed that two domains, acidic and bHLH, were significantly associated with petal color phenotypes, then five amino acid sites in these two domains (three at acidic domain and two at bHLH domain) were selected for further multiple regression analyses using GLM. The significance of association was tested using F statistic at levels of $P<0.1$, $P<0.05$, and $P<0.01$.

Reconstruction of Gene Genealogies

Traditional phylogenetic analysis approaches (e.g. parsimony, maximum likelihood, and neighbor-joining) focus on the estimation of interspecific phylogenies, and these methods require a relatively large number of variable characters to reconstruct relationships

(Huelsenbeck & Hillis 1993). Intraspecific haplotypes, however, provide fewer variable and informative sites to resolve relationships. Moreover, phylogenetic relationships at the intraspecific level are affected by many processes (e.g. effective population size, allele frequency, gene flow, and recombination) that are generally ignored in interspecific phylogenetic reconstruction (Templeton et al. 1992). Therefore, reconstruction of intraspecific phylogenies at the population level by traditional methods will often lead to erroneous or unresolved conclusions (Clement et al. 2000). Considering hybridization, introgression, gene flow, and recombination occurring within the dwarf dogwoods complex, we estimated the gene genealogy using a haplotype network methodology (Templeton et al. 1992) that takes into account the population level phenomena. Haplotypes were collapsed from DNA sequences using the computer program COLLAPSE 1.1 (Posada 2003). The haplotype file generated from COLLAPSE was used to estimate the gene genealogy. Gaps were treated as fifth states. Genealogical relationships were estimated using 95% statistical parsimony support criterion (Templeton et al. 1992) as implemented in the computer program TCS 1.13 (Clement et al. 2000).

Results

Sequence data

The anthocyanin regulatory gene sequences generated for the 47 samples of dwarf dogwoods varied from 4023 to 4040 base pairs in length except for two with incomplete sequences (samples #40-1 and #35-1). The sequences contain eight exons, seven introns, and flanking regions. The length of the coding region ranges 1863, 1875, or 1878 base pairs, and

seven introns have a total of 1864 or 1867 base pairs (Fig. 2). In addition to nucleotide substitutions, a total of four indels were found, of which two were identified in exon VI, and one each was detected in intron 1 and 2, respectively. For exon indels, three base pairs (TAT) are unique to hybrids (Table 4); and the other 12-bp indel serves as the identification marker between *C. canadensis* and *C. suecica*. The presence of the 12 base pair sequence is characteristic of specimen of *C. suecica*; the absence of the 12 nucleotides is characteristic of specimen of *C. canadensis* (Table 4). The hybrids can be identified as heterozygous for indels (Table 4). For intron indels, sample # 27-1, 27-2, 43-1, 43-2, and 94-388 lack of one repeat of 'AT' at intron 1 and a single additional 'G' was found for sample #5-3, 6-1, and 15-1 at intron 2 (Fig.2). The data matrix contains 186 variable sites (4.60%) and 105 (2.60%) parsimony informative sites within the dwarf dogwoods, of which 83 variable sites and 45 informative sites come from introns; and 88 variable sites and 51 informative sites from exons.

Nucleotide diversity and excess low-frequency polymorphism

The level of sequence variation observed in the three groups is highly different. Sequence analyses indicated that the 12 samples of group 'CC' (*C. canadensis*) contain 39 segregating sites and each has a distinct haplotype. This group has the lowest sequence diversity among all π values (including all sites and different types of sites) and lowest θ_w value, suggesting that substantial gene flow occurs within this group. Silent-site nucleotide diversity (π_{silent}) in this group is 0.00253, which is comparable to the estimated population mutation parameter ($\theta_w = 0.00341$) (Table 5). Among 40 mutation sites across the entire sequence, twenty-one are from protein coding regions. For polymorphic sites in exons,

sixteen are nonsynonymous mutations and five are silent mutations (Table 6). In contrast, group of 'CS' (*C. suecica*) has the highest sequence variation and diversity in terms of π and θ_w values, which are approximately three times the values for group 'CC' in all tests. Silent sequence diversity (π_{silent}) is 0.00763, also comparable to the estimated population mutation parameter (θ_w = 0.00881) (Table 5). The seven samples of 'CS' form seven haplotypes with 82 segregating sites. Forty-seven of 82 segregating sites are from coding regions. For polymorphic sites in coding regions, thirty-six are replacement mutations, and 11 are synonymous mutation (Table 6). Group 'CH' has intermediate values of π and θ_w between those for group 'CC' and group 'CS'. Silent sequence diversity (π_{silent}) is 0.00409, which is comparable to the estimated population mutation parameter (θ_w = 0.00565) (Table 5). A total of 82 mutation sites were found in group 'CH", of which 40 are from coding regions. Thirty-two of the 40 sites in coding regions are replacement mutations, and eight are silent mutations (Table 6). The sequence divergence test shows that group 'CC' and 'CS' are most divergent; group 'CH' is closer to 'CC' than to 'CS' (Table 7). All neutrality hypothesis tests examined in this study yielded negative D values for both Tajima's and Fu & Li's tests, suggesting non-neutrality of the gene evolution in the complex. Furthermore, coalescent simulations found a significant difference from the neutral-equilibrium model expectation within the entire species complex and within each of the three groups (Table 5). These results indicate the presence of an excess of low-frequency polymorphisms within the dwarf dogwoods and within each of the three groups, suggesting positive selection on the coding regions.

In all groups, the majority of silent site variation comes from mutations in introns rather than at synonymous sites. In group 'CC', only 5 of 24 silent mutations are from exons,

in ‘CS’, 8 out of 50, and in group ‘CH’, 11 out of 47 silent sites (Table 6). This suggests a higher rate of neutral mutation in the intron regions. In coding regions, there are more nonsynonymous mutations than synonymous mutations (16 versus 5 in ‘CC’; 32 versus 8 in ‘CH’; and 36 versus 11 in ‘CS’) (Tables 6).

Associations between amino acid sequence variation and petal colorization

The ANOVA table indicates that there is a significant correlation between protein sequence variation and petal color with *P* value less than 0.0001 (Table 9). Comparison among domains indicates that two of the four domains, acidic (at P=0.05 level) and bHLH (at P=0.1 level), are significantly associated with petal color (Table 10). Individual site tests detected three sites significantly correlated with petal color. Two of these sites, site #436 and site #464, occur in the bHLH domain, and one site (site #307) is found in the acidic domain (Table 11). The changes are a substitution between alanine (A) and serine (S) at site 307, serine (S) and leucine (L) at site 436, and valine (V) and isoleucine (I) at site 464. Haplotypes at these three sites show that ASV is associated with wide petals, SSI is associated with bicolor petals, and SLI is associated with purple color. Furthermore, an indel of four amino acid sites is strongly correlated with petal color. In white petal samples, these four amino acids are absent; in bicolor petal and purple petal samples, these four amino acids are present.

Genealogy of the myc-like anthocyanin regulatory gene

Forty-seven samples were collapsed into 46 haplotypes (haplotype 25-2 included two samples: 25-2 and 28-1). The *myc*-like anthocyanin regulatory gene genealogy estimation

using TCS revealed that two haplotypes, 43-2 and 94-388 (*C. suecica*), of 46 haplotypes were unconnected due to their high sequence divergence from other haplotypes. Figure 4 shows the 95% statistical parsimony based haplotype network for the remaining 44 haplotypes. In this haplotype network, two separate clades were revealed. The first clade includes five haplotypes from *C. suecica* or *C. canadensis*<*C. suecica* (lineage 'D'), bearing completely purple petals. This clade has many fixed sequence variations that are not found in other haplotypes. The second clade contains *C. canadensis* and hybrids mostly with white colors or bicolor petals (Fig. 4), and has no fixed sequence differences between white petal haplotypes and bicolor petal haplotypes. However, 10 of 12 white petal haplotypes form a subgroup with a few fixed polymorphisms (Fig. 4). Furthermore, three regions connections among haplotypes were found in TCS network. The first nonlinear connection involves five white petal haplotypes (C18-3, C19-1, C15-1, C17-1, and C40-1) (Fig. 4), indicating gene flow and recombination among *C. canadensis*. This result is congruent with the sequence diversity tests. The nonlinear connection was also found for several hybrids and *C. canadensis* (e.g. U3-1, U36-3, U34-1, U25-2, and U23-1) (Fig. 4), suggesting that gene flow and recombination occurs among hybrids and *C. canadensis*. The recombination event suggested by the TCS network is congruent with estimates of recombination rates (Table 8).

Discussion

Accelerated and non-neutral evolution of the myc-like anthocyanin regulatory gene evolution in dwarf dogwoods

Regulatory genes play an important role in the evolution of plants and animals (reviews by Purugganan 1998, 2000; Levine & Tjian 2003; Papp et al. 2003). Theoretical and empirical studies have demonstrated that changes in regulatory genes can cause dramatic shifts in morphology and ecological, physiological activities (Wilson 1975; King & Wilson 1975; Carroll 1995; Palopoli & Patel 1996; Doebley & Lukens 1998; Purugganan 1998; Simpson 2002; Kellogg 2002). Detailed investigation of regulatory loci both within populations and between closely related species is likely required to reach a comprehensive understanding of the evolutionary dynamics of regulatory genes. Estimates of sequence variation and gene genealogies for regulatory genes across species boundaries will allow us to connect micro-evolutionary forces within species to macro-evolutionary changes between species. Recent studies in *Arabidopsis*, *Zea mays*, *Brassica*, and *Drosophila* detected that regulatory genes can harbor significant variation responsible for adaptive morphological evolution (see review by Purugganan 2000). However, the complete roles of regulatory genes in evolutionary diversification as well as the links between phenotypic variation and sequence variation still remain mysteries. Regulatory loci could maintain low molecular diversity, yet under positive selection. For example, the *B_0Cal* gene exhibits low sequence variation in *Brassica*, but positive selection at this regulatory locus was detected (Purugganan et al. 2000, 2001). On the other hand, accelerated protein polymorphisms have been documented in both plants and animals. For example, the elevated rates of gene evolution associated with reproductive morphological diversification following hybridization and polyploidization have been demonstrated in catostomid fish species using multilocus isozymes (Ferris & Whitt 1979). In *Arabidopsis*, three floral developmental genes (*CAL*,

AP3, and *PI*) show accelerated protein polymorphism, which was linked to positive selective pressure (Riechmann & Meyerowitz 1997; Purugganan & Suddith 1998, 1999).

The *myc*-like anthocyanin regulatory genes in dwarf dogwoods add another example of elevated sequence evolution in regulatory genes. Both excess of low-frequency polymorphisms and excess of replacement mutations were detected within the dwarf dogwoods species complex. There are two possible forces acting on this rapid sequence evolution. First, positive selection may be the major forces to promote an excess of low-frequency polymorphism and replacement site polymorphism in the complex. Secondly, recent studies in *Arabidopsis*, wild barley, and the Hawaiian silversword alliance suggest rapid population expansion as another mechanism that may explain an excess of low-frequency polymorphism and replacement site polymorphisms in plants (Cummings & Clegg 1998; Purugannan & Suddith 1998; Lawton-Rauh et al. 2003). Previous studies (Bain & Denford 1979; Murrell 1994) suggested that in the Pleistocene, previously allopatric *C. canadensis* and *C. suecica* expanded their geographic distribution and met in the pacific regions of Northwestern of America, resulting in hybridization and polyploidization. The complex probably expanded their distribution again after glaciations and resulted in the current wide distribution in the circumboreal region. Thus, it is possible that a recent range expansion of the group may also contribute to the observed excess of low-frequency polymorphisms and excess of replacement mutations.

Link between color of petals and gene evolution

Our sequence analyses and ANOVA tests reveal a link between the color of petals and amino acid sequence variation. Further, an ANOVA test indicates that changes of three

amino acid sites in the acidic domain and bHLH domain of the gene are significantly associated with the color change of petals. The function of the bHLH domain is believed to involve DNA-binding, as well as subunit dimerization activity of the *myc*-like/*R* protein (Murre et al. 1989, 1994; Atchley & Fitch 1997; Atchley et al. 2000). Thus, mutation in the bHLH domain would critically affect the function of the protein. The bHLH domain includes 56 amino acids in dwarf dogwoods, comprising basic region, two helix regions, and one loop region. The basic region includes basic residues that permit protein binding to a consensus hexanucleotide E-box. The Helix-Loop-Helix region contains hydrophobic residues that allow formation homo- or heterodimers (Atchley et al. 1999). In dwarf dogwoods, the basic region extends from site 431 to site 443, helix 1 extends from site 444 to site 458; and helix 2 extends from site 472 to site 486. Site 436, located in the basic region, shows a substitution between serine (S) and leucine (L). Leucine is mainly found in samples with purple petals. Leucine, a hydrophobic residue, is required for the formation dimer. Serine, in contrast, has the hydrophilic side chain that may be required for binding in the basic region. Thus, substitutions between leucine and serine may cause improper binding. Site 464 is located in the loop region which exists to provide spacing required for dimerization. Previous studies have indicated that its amino acid content is not very important for the function and evolution of this protein (Atchley et al. 2000). Therefore, substitution between isoleucine and valine, which have similar chemical features, are likely to have no obvious effect on function of bHLH domain. Our analyses also identified a third amino acid (site 307, alanine *vs* serine) in the acidic domain associated with petal color. At this site, alanine, a hydrophobic residue, is found only in samples with white petals. In contrast, serine, with a hydrophilic side chain, is found in samples with purple or partially purple petals. The acidic domain plays a role in

transactivation which requires an amino acid with hydrophilic side chain (Gong et al 1999). Previous comparison between MYC-RP/GP in *Perilla* and *Delia* in *Antirrhinum majus* found that *Delia* exhibited higher transactivation activity than MYC-RP/GP (Gong et al. 1999). Therefore, amino acid substitutions in this region may be the main reason for the different transactivation activities observed.

Two of samples (sample #22-1 and #23-1) have red petals but have bicolor petal genotype. Both other morphological features and sequence analyses reveal that these two samples are hybrids. In addition to genetic factors, chemical and environmental characteristics (e.g. light intensity, vacuolar pH, and cell shape) can also be involved in the production of pigmentation that will alter color of flower (Mol et al. 1998). For example, in *Petunia*, several mutated loci that increase the pH of petal extracts, cause blueing of the flower (Mol et al. 1998). Intensive light can initiate the red/purple color of flowers that has studied in vivo in transgenic plants (Ray et al. 2003). Our red samples were found on the top of a hill with great light intensity, thus, the red color may be a result induced by environmental factors.

Gene flow and gene recombination within dwarf dogwood species complex

Challenges remain in determining the origins of hybrids and polyploids. Certainly, morphological intermediacy has been a major clue in identifying hybrids and polyploids. However, morphology may be misleading (Rieseberg 1995). Molecular markers have been demonstrated to provide a powerful tool for elucidating the parental origins of hybrids and polyploids (Soltis et al. 1992). The dwarf dogwood complex was previously investigated in morphology, ecology, and biogeography (Bain & Denford 1979; Murrell 1994). Murrell

(1994) identified five lineages based on morphology. Lineage A, C, and E were recognized as distinct species, *C. canadensis*, *C. unalaschkensis*, and *C. suecica*, respectively. *Cornus unalaschkensis* is a tetraploid considered to have derived from hybridization and polyploidization between two diploid species (*C. canadensis* and *C. suecica*) (Dermen 1931; Taylor & Brockman 1966; Clay & Nath 1971; Bain & Denford 1979). Lineage B and D were considered to be products of hybridization between *C. canadensis* and *C. suecica* and subsequent backcrossing with *C. canadensis* and *C. suecica*, respectively. We detected a 12-bp indel occurring in exon VI that distinguishes the purple and white petal color phenotypes corresponding to two diploid species (Table 4 and Fig. 2). In the bicolor group, each individual is a heterozygous with one allele containing the 12-bp and another allele without the 12 base pairs.

Morphological characters between *C. canadensis* and hybrids are highly consecutive and not easily categorized. Particularly, lineage B (*C. canadensis*>*C. suecica*) is the most problematic lineage in the dwarf dogwoods complex. This group is hard to differentiate from *C. canadensis* due to its high morphological similarity to *C. canadensis* (Murrell 1994). Corresponding with sequence variation evidence, this morphological similarity could be caused by gene flow and chromosomal recombination, causing introgression to *C. canadensis*. Our data from anthocyanin regulatory gene indicate more gene flow between *C. canadensis* and *C. canadensis*>*C. suecica* in the Pacific Northwest. The intermediate forms with bicolor petals are genetically more similar to *C. canadensis*. The gene recombination was also suggested for all three groups. However, these factors, e.g. gene flow and recombination, can also prevent speciation and inhibit the population subdivision necessary for reproductive isolation among populations. Theoretical models and empirical studies have

demonstrated that gene recombination can oppose speciation (see review by Ortiz-Barrientos et al. 2002), while reductions in recombination and/or evolved genetic correlations could possibly contribute to species formation and persistence (Ortiz-Barrientos et al. 2002). On the other hand, hybridization following introgression may lead to the transfer of traits from one taxon into another, allowing for range expansion of the introgressed form (Lewontin & Birch 1966). Most of the population occurring in the Pacific Northwest are bicolor in petals resulted from hybridization. Thus, the bicolor form of the dwarf dogwoods may have an advantage to expand its range in the Pacific Northwest Mountains.

In conclusion, our data show that the *myc*-like anthocyanin regulatory gene evolves at an accelerated rate in dwarf dogwoods, as evidenced by an excess of low-frequency polymorphisms and replacement sites. This rapid rate may have been caused by both positive selection and/or recent, rapid population expansion. The few examined *C. suecica* populations appear to have high molecular diversity with many fixed mutations, suggesting isolation of these plants from the rest of populations sampled. *C. canadensis* and hybrids appear to have extensive gene flow, suggesting little genetic isolation between them, and/or recombination within and between *C. canadensis* and hybrids. We also detect strong correlation between petal colors and three functionally important amino acid sites.

It remains unclear to what extant genome reconstructing as a result of polyploidization and what molecular speciation pattern occurring within the dwarf dogwoods species complex, given that this study was restricted sampling from Pacific regions. Therefore, more studies (with complete samples across North of American and multiple markers) will be needed to unravel the precise relationships between polyploidization, the diversification of

homoeologous regulatory gene copies, and the adaptive radiation of the dwarf dogwood complex.

Acknowledgments

The authors thank the following people for providing different kinds of help: Brian Cassel for assistance with sequencing; Jim Qi for sample collection in the Alaska field trip; Nina Gardner for DNA extraction and morphological identification; members of the Xiang lab for a variety of help and discussion. This study is supported by Faculty Research Grant money from Idaho State University and North Carolina State University and NSF grant DEB-0129069 to Q-Y.X.; Karling Graduate Student Research Award from Botanical Society of America and Deep Gene Travel Award from Deep Gene Network of NSF to C.F.

References

Arnold ML (1997) *Natural Hybridization and Evolution*. Oxford University Press, Oxfoird, UK.

Arnold ML, Kentner EK, Johnston JA, Cornman S, Bouck AC (2001) Natural hybridization and fitness. *Taxon*, 50, 93-104.

Atchley WR, Fitch WM (1997) A natural classification of the basic helix-loop-helix class of transcription factors. *Proceedings of National Academy of Sciences USA*, 94, 5172-5176.

Atchley WR, Terhalle W, Dress A (1999) Positional dependence, cliques, and predictive motifs in the bHLH protein domain. *Journal of Molecular Evolution*, 48, 501-516.

Atchley, WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW (2000) Correlation among amino acid sites in bHLH protein domains: an information theoretic analysis. *Molecular Biology and Evolution*, 17, 164-178.

Bain JF, Denford KE (1979) The herbaceous members of genus *Cornus* in NW North America. *Botanical Notiser*, 132, 121-129.

Barrier M, Baldwin BG, Robichaux RH, Purugganan MD (1999) Interspecific hybrid ancestry of a plant adaptive radiation: allopolyploidy of Hawaiian silversword alliance (Asteraceae) inferred from floral homeotic gene duplications. *Molecular Biology and Evolution*, 16, 1105-1113.

Barrier M, Robichaux RH, Purugganan MD (2001) Accelerated regulatory gene evolution in an adaptive radiation. *Proceedings of National Academy of Sciences USA*, 98, 10208-10213.

Brochmann C, Soltis PS, Soltis DE (1992) *American Journal of Botany*, 79, 673-688.

Burke JM, Arnold, ML (2001) Genetics and fitness of hybrids. *Annual Review of Genetics*, 35, 31-52.

Carroll SB (1995) Homeotic genes and the evolution of arthropods and chordates. *Naturalist*, 376, 479-485.

Clay SN, Hath J (1971) Cytogenetics of some species of *Cornus*. *Cytologia*, 36, 716-730.

Clement M, Posada D, Crandall KA (2000) TCS: a computer program to estimate gene genealogies. *Molecular Ecology*, 9, 1657-1659.

Consonni G, Viotti A, Dellaporta SL, Tonelli C (1992) cDNA nucleotide sequence of *Sn*, a regulatory gene in maize. *Nucleic Acids Research*, 20, 373.

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

Cronn RC, Small RL, Wendel J F (1999) Duplicated genes evolve independently after polyploid formation in cotton. *Proceedings of National Academy of Sciences USA*, 96, 14406-14411.

Cummings MP, Clegg MT (1998) Nucleotide sequence diversity at the *Adh1* locus in wild barley, an evaluation of the background selection hypothesis. *Proceedings of National Academy of Sciences USA*, 95, 5673-5642.

Dermen H (1932) Cytological studies of *Cornus*. *Journal of the Arnold Arboretum*, 13, 410-417.

De-Vetten N, Quattrocchio F, Mol J, Koes R (1997) The *an1* 1 locus controlling flower pigmentation in *Petunia* encodes a novel WD-repeat protein conserved in yeast, plants, and animals. *Genes and Development*, 11, 1422-1434.

Dickinson WJ (1988) On the architecture of regulatory systems: evolutionary insights and implications. *BioEssays*, 8, 204-208.

Doebley, J (1993) Genetics, development and plant evolution. *Current Opinion in Genetics and Development*, 3, 865-872.

Doebley J, Lukens L (1998) Transcriptional regulators and the evolution of plant form. *Plan Cell*, 10, 1075-1082.

Ferris S, Whitt G (1979) Evolution of the differential regulation of duplicate genes after polyploidization. *Journal of Molecular Evolution*, 12, 367-317.

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. *Genetics,* 133, 693-709.

Goldsbourough, AP, Tong Y, Yoder JI (1996) *Lc* as a non-destructive visual reporter and transposition marker gene for tomato. *Plant Journal*, 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 Molecular Biology*, 41, 33-44.

Goodrich J, Carpenter R, Coen ES (1992) A common gene regulates pigmentation pattern in diverse plant species. *Cell*, 68, 955-964.

Grant V (1981) *Plant Speciation*. Columbia University Press, New York, USA.

Hey J, Wakeley J (1997) A coalescent estimator of the population recombination rate. *Genetics*, 145, 833-846.

Hu J, Anderson B, Wessler SR (1996) Isolation and characterization of rice *R* genes: evidence for distinct evolutionary paths in rice and maize. *Genetics*, 142, 1021-1031.

Huelsenbeck JP, Hillis DM (1993) Success of phylogenic methods in the four-taxon case. *Systematic Biology*, 42, 247-264.

Kellogg EA (2002) Root hairs, trichomes and the evolution of duplicate genes. *Trends in Plant Science*, 6, 550-552.

King JL, Wilson AC (1975) Evolution at two levels in humans and chimpanzees. *Science*, 188, 107-116.

Lamarck J (1786) *Encyclopedie Methodique*, Vol. 2. Liege: Chez Plomteux.

Lawton-Rauh A, Robichaux RH, Purugganan MD (2003) Patterns of nucleotide variation in homoeologous regulatory genes in the allotetraploid Hawaiian silversword alliance (Asteraceae). *Molecular Ecology*, 12, 1301-1313.

Ledebour CF (1844) *Flora Rossica*, vol. 2. Stuttgart: Schweizerbart.

Levine M, Tjian R (2003) Transcription regulation and animal diversity. *Nature* 424:147-151.

Lewonton RC, Birch LC (1966) Hybridization as a source of variation for adaptation to new environments. *Evolution*, 20, 3150336.

Lioyd AM, Walbot V, Davis RW (1992) *Arabidopsis* and *Nicotiana* anthocyanin production activated by maize regulators *R* and *C1*. *Science*, 258, 1773-1775.

Liu YG, Whittier RF (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 YG, Mitsukawa N, Oosumi T, Whittier RF (1995) Efficient isolation and mapping of *Arabidopsis thaliana* T-DNA insert junctions by thermal asymmetric interlaced PCR. *Plant Journal*, 8, 457-463.

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

Martin C, Prescott A, Mackay S, Barlett J, Vrijlandt E (1991) Control of anthocyanin biosynthesis in flower of *Antirrhinum majus*. *Plant Journal*, 1, 37-49.

Masterson J (1994) Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. *Science*, 264, 421-424.

Mol J, Grotewold E, Koes R (1998) How genes paint flowers and seeds. *Trends in Plant Science*, 3, 212-217.

Morgan DR, Soltis DE (1993) Phylogenetic relationships among members of Saxifragaceae sensu lato based on *rbcL* sequence data. *Annals of the Missouri Botanical Garden*, 80, 631-660.

Murre C, McCaw PS, 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 MA, Engel I, Furnari BA, Massari ME, Mathews JR, Quong MW, Rivera RR, Stuiver MH (1994) Structure and function of helix-loop-helix proteins. *Biochimica Et Biophysica Acta*, 1218, 129-135.

Murrell ZE (1994) Dwarf dogwoods: intermediacy and the morphological landscape. *Systematic Botany*, 19, 539-556.

Nei M (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York, USA

Ortiz-Barrientos D, Reiland J, Hey J, Noor MAF (2002) Recombination and the divergence of hybridizating species. *Genetica*, 116, 167-178.

Otto SP, Whitton J (2000) Polyploid incidence and evolution. *Annual Review of Genetics*, 34, 401-437.

Palopoli MF, Patel N (1996) Neo-Darwinian developmental evolution- can we bridge the gap between pattern and process? *Current Opinion in Genetics and Development*, 6, 502-508.

Papp B, Pál C, Hurst LD (2003) Evolution of *cis*-regulatory elements in duplicated genes of yeast. *Trends in Genetics*, 19, 417-422.

Posada D (2003) COLAPSE 1.1 (MAC). Crandall lab computer programs. http://inbio.byu.edu/Faculty/kac/crandall_lab/programs.htm.

Purugganan MD (1998) The molecular evolution of development. *BioEssays*, 20, 700-711.

Purugganan MD (2000) The molecular population genetics of regulatory genes. *Molecular Ecology*, 9, 1451-1461.

Purugganan MD, Suddith JI (1998) Molecular population genetics of the *Arabidopsis CAULIFLOWER* regulatory gene: nonneutral evolution and naturally occurring variation in floral homeotic function. *Proceedings of National Academy of Sciences USA*, 95, 8130-8134.

Purugganan D, Suddith JI (1999) Molecular Population Genetics of Floral Homeotic Loci: Departures from the equilibrium-neutral model at the *APETALA3* and *PISTILLATA* genes of *Arabidopsis thaliana*. *Genetics* 1999, 151, 839-848.

Purugganan MD, Boyles AL, Suddith JI (2000) Variation and selection at the *CAULIFLOWER* floral homeotic gene accompanying the evolution of domesticated *Brassica oleracea*, *Genetics*, 155, 855-862.

Quattrocchio F, Wing JF, Leppen HTC, Mol JNM, Koes RE (1993) Regulatory genes controlling anthocyanin pigmentation are functionally conversed among plant species and have distinct sets of target genes. *Plant Cell*, 5, 1497-1512.

Quattrocchio F, Wing JF, Woude KVD, Mol JNM, Koes RE (1998) Analysis of bHLH and MYB domain proteins: species-specific regulatory difference are caused by divergent evolution of target anthocyanin genes. *Plant Journal*, 13, 475-488.

Radicella PD, Turks D, Chandler VL (1991) Cloning and nucleotide sequence of a cDNA encoding B-Peru, a regulatory protein of the anthocyanin pathway in maize. *Plant Molecular Biology*, 17, 127-130.

Ray H, Yu M, Auser P *et al*. (2003) Expression of anthocyanins and proantocyanins after transformation of alfalfa with maize *Lc*. *Plant Physiology*, 132, 1448-1463.

Riechmann JL, Meyerowitz EM (1997) MADS domain proteins in plant development. *Biological Chemistry*, 378, 1079-1101.

Rieseberg LH (1995) The role of hybridization in evolution: old wine in new skin. *American Journal of Botany*, 82, 944-953.

Rieseberg LH (1997) Hybrid origins of plant species. *Annual Review of Ecology and systematics*, 28, 359-389.

Rieseberg LH, Wendel JF (1993) Introgression and its consequences in plants. In: *Hybrid Zones and the Evolutionary Process* (ed Harrison RG), pp. 70-109. Oxford Univ. Press, New York.

Rozas J, Sánchez-Delbarrio JC, Messeguer X, Rozas R (2003) DNASP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*, 19, (in press).

SAS Institute Inc. (1999) *SAS/STAT @ User's Guide*, Version 8. Cary, NC.

Simpson P (2002) Evolution of development in closely related species of flies and worms. *Nature Reviews Genetics*, 3, 907-917.

Soltis PS, Doyle JJ, Soltis DE (1992) Molecular data and polyploid evolution in plants. In: *Molecular Systematics of Plants* (eds Soltis PS, Soltis DE, Doyle JJ), pp.177-201. Chapman & Hall, New York.

Soltis PS, Gitzendanner MA (1999) Molecular systematics and the conservation of rare species. *Conservation Biology*, 13, 471-483.

Soltis DE, Soltis PS (1997) Phylogenetic relationships among Saxifragaceae sensu lato: a comparison of topologies based in 18S rDNA and *rbc*L sequences. *American Journal Botany*, 84, 504-522.

Soltis DE, Soltis PS (1999) Polyploidy: recurrent formation and genome evolution. *Trends in Ecology and Evolution*, 14, 348-352.

Soltis PS, Plunkett G, Novak S, Soltis DE (1995) Genetic variation in *Tragopogon* species: additional origins of the allotetraploids *T. mirus* and *T. miscellus* (Compositae). *American Journal of Botany*, 82, 1329-1341.

Song K, Lu P, Tang K, Osborn TC (1995) Rapid genome change in synthetic polyploids of Brassica and its implications for polyploid evolution. *Proceedings of National Academy of Sciences USA*, 92, 7719-7723.

Stebbins GL (1971) *Chromosomal Evolution in Higher Plants*. Edward Arnold, London.

Tajima F (1989) Statistical methods for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123, 595-595.

Taylor RL, Brockman RP (1966) Chromosome numbers of some western Canadian plants. *Canadian Journal of Botany*, 44, 1093-1103.

Templeton AR, Crandall KA, Sing CF (1992) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III Cladogram estimation. *Genetics*, 132, 619-633.

Watterson G (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7, 256-276.

Wendel JF (2000) Genome evolution in polyploids. *Plant Molecular Biology*, 42, 225-249.

Wendel JF, Schnabel A, Seelanan T (1995) Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*). *Proceedings of National Academy of Sciences USA* 92: 280-284.

Wilson AC (1975) Evolutionary importance of gene regulation. *Stadler Symposium*, 7, 117-134. Columbia, Mo.: University of Missouri.

Xiang, QY, Soltis DE, Soltis PS (1998) Phylogenetic relationships of Cornaceae and close relatives inferred from *matK* and *rbcL* sequences. *American Journal of Botany*, 85. 285-297.

Table 1. The dwarf dogwood samples used in this study. Lineage identification is determined by the morphological criteria of Murrell (1994)

Lineages	Total number of populations (localities)	Samples: population (locality) #-individual #	Color of petals	Collection Localities (see Fig1.)
A (*C. canadensis*)	9	6-1, 7-1, 8-1, 17-2, 18-3, 21-2, 32-2, 33-2, 40-1	White	AK, BC, YK
B (*C. canadensis* > = *C. suecica*)	6	9-1, 14-1, 15-1, 16-1, 19-1, 20-1	White except # 9-1 with bicolor	AK, BC, YK
C (*C. unalaschkensis*)	22	1-11, 2-6, 3-1, 4-1, 5-3, 10-1, 11-1, 12-1, 13-1, 22-1, 23-1, 24-1, 25-2, , 26-1, 28-1, 31-1, 34-1, 35-3, 36-3, 37-1, 38-2, 39-4	Bicolor with white range from1/3 to 2/3, except #22-2 and 23-1 with red petals	ID, WA, OR, BC, AK
D (*C. canadensis* =< *C. suecica*)	4	29-1, 29-2, 30-1, 41-1, 42-1	Purple (29-1 and 29-2) or mostly purple (others)	AK
E (*C. suecica*)	3	27-1, 27-2, 43-1, 43-2, 94-388	Purple	AK, Norway

Table 2. The locus specific primers used for PCR amplifying and sequencing the *myc*-like anthocyanin regulatory gene

Name of primer	Length (bp)	Sequence (5'-3')	Tm	Location
F0A	21	TCACTGAGTGGGTGTCTTAAG	55.0	5'-Flank
F2A	22	TTTATGAGTCCCTTGYGGTCAC	58.5	Exon II
F2A1	25	GTTCAGGCGGTCGAATTCAATGCCG	64.2	Exon II
R2A	22	GTGACCRCAAGGGACTCATAAA	57.2	Exon II
R3A	20	AGAGCGACTGAATATTTTGC	51.9	Exon III
R3'	20	CCSAGCTCAAYYACWCCTCC	57.8	Exon V
F4A	20	GCGATATTGCCATTTGTCTG	53.2	Intron 4
F4A3	20	TATGGTAGAAGGCTTAAATG	47.9	Exon VI
R4A	21	CATTTATGGAAGTAAGGTCCC	51.6	Exon VI
F6A	19	CTGACCTCGTTGGACCTTC	55.6	Exon VI
R7A	19	CAAGCAACAAGCGCTCCCT	59.5	Exon VII
F7A2	19	GAGCTGGAGATCAACCTCG	55.4	Exon VII
R8A3	21	AAGTGCTTGTATGATCATCCC	53.8	Exon VIII
R9A	21	CTATCCACAAGAAACACYTGC	53.8	3'-Flank

*F: forward primers; R-reverse primers

Table 3. The specific primers and arbitrary degenerated primers used in TAIL-PCR

Name	Length (bp)	Sequences (5'-3')	Tm	Location	Notes
AD1	15	NTCGASTWTSGWGTT	46.0	Unknown	Liu et al, 1995
AD2	16	NGTCGASWGANAWGAA	46.8	Unknown	Liu et al, 1995
AD3	16	WGTGNAGWANCANAGA	34.8	Unknown	Liu et al, 1995
R1A1	25	CTTATGCCAAAAGACATGTTCTTGA	58.1	Intron 1	Used for primary reaction to amplify the 5'-flanking region
R1A2	25	TTGAATTCTAACTTACATCTACATG	54.8	Intron 1	Used for secondary reaction to amplify the 5'-flanking region
R1A3	20	CATGGCACCCAATTATTATT	51.2	Intron 1	Used for tertiary reaction to amplify the 5'-flanking region
F5A1	22	GGCACAACCGACCTTTTCTTA	57.4	Exon VI	Used for primary reaction to amplify the 3'-flanking region
F5A2	22	TCGGTCCTTGGATCATTGATCC	57.9	Exon VI	Used for secondary reaction to amplify the 3'-flanking region
F5A	20	TGATCCCTTCCACTAGCAAG	54.9	Exon 6	Used for tertiary reaction to amplify the 3'-flanking region

*F: forward primers; R-reverse primers

Table 4. The distribution of two exon indels (12-bp and 3-bp) in 47 samples of the dwarf dogwoods

Nucleotide sequence	Samples without sequence	Samples with sequence	Samples with heterozygote
CTTGATG CGGCT	6-1, 7-1, 8-1, 14-1, 15-1, 16-1, 17-2, 18-3, 19-1, 21-2, 32-2, 33-2, 40-1	27-1, 27-2, 43-1, 43-2, 94-388	1-11, 2-6, 3-1, 4-1, 5-3, 9-1, 10-1, 11-1, 12-1, 13-1, 20-1, 22-1, 23-1, 24-1, 25-2, 26-1, 28-1, 29-1, 29-2, 30-1, 31-1, 34-1, 35-3, 36-3, 37-1, 38-2, 39-4, 41-1, 42-1
TAT	4-1, 6-1, 7-1, 8-1, 11-1, 12-1, 13-1, 14-1, 15-1, 16-1, 18-3, 19-1, 21-2, 22-1, 24-1, 25-2, 27-1, 27-2, 28-1, 29-1, 29-2, 30-1, 31-1, 32-2, 33-2, 41-1, 42-1, 43-1, 43-2, 94-388		1-11, 2-6, 3-1, 5-3, 9-1, 10-1, 17-2, 20-1, 23-1, 26-1, 34-1, 35-1, 36-3, 37-1, 38-2, 39-4, 41-1

Table 5. Molecular diversity in the *myc*-like anthocyanin regulatory gene within the dwarf dogwoods

Group	Length	*n*	N_{hap}	S	π_{all}	π_{silent}	$\pi_{noncode}$	π_{synon}	π_{nonsyn}	θ_w ($4N\mu$)	Tajima's D	Fu and Li's D
All	3928	45	45	165	0.00643	0.00680	0.00703	0.00569	0.00586	0.00997	-1.40476	-2.02896
											P=0.036**	*P*=0.015**
CC	3928	12	12	39	0.00253	0.00217	0.00193	0.00327	0.00232	0.00341	-1.58211	-1.82960
											P=0.005***	*P*=0.004***
CS	3928	7	7	82	0.00763	0.00692	0.00609	0.01082	0.00885	0.00881	-0.78645	-0.77685
											P=0.042**	*P*=0.038**
CH	3928	26	26	82	0.00409	0.00423	0.00389	0.00584	0.00387	0.00565	-1.07879	-1.70403
											P=0.0135**	*P*=0.0195**

All: all samples grouped into one group; CC: *C. canadensis*; CS: *C. suecica*; CH: hybrid; *n*: the number of individuals included; N_{hap}: the number of observed haplotypes; S: the number of observed segregating sites. π_{all}: nucleotide diversity at all sites; π_{silent}: nucleotide diversity at silent sites (synonymous and noncoding regions); $\pi_{noncode}$: nucleotide diversity at noncoding regions (franking region and introns); π_{synon}: nucleotide diversity at synonymous sites; π_{nonsyn}: nucleotide diversity at nonsynonymous sites; θ_w: Watterson's estimate of mutation parameter. * Significant at the $P<0.1$ level; ** significant at the $P<0.05$ level; ***significant at the $P<0.01$ level.

Table 6. Summary of sequence diversity in ‘CC’, ‘CH’, and ‘CS’

	Region	Total sites	Silent sites			Nonsynonymous sites			Silent mutation			Replacement mutation		
			CC	CH	CS	CC	CH	CS	CC	CH	CS	CC	CH	CS
5' Flanking	1-61	61	61	61	61				0	2	0			
Exon 1	62-189	128	26.67	26.87	26.33	101.33	101.13	101.67	0	0	0	0	1	0
Intron 1	190-744	554	554	555	553				4	6	14			
Exon 2	745-999	255	56.89	55.83	55.83	196.11	197.17	197.17	1	3	2	2	2	1
Intron 2	1000-1535	531	531	531	530				4	8	6			
Exon 3	1536-1632	97	24.33	24.19	24.19	72.67	72.81	72.81	0	0	0	0	2	2
Intron 3	1633-1841	199	199	199	199				3	7	7			
Exon 4	1842-1856	15	4.00	4.00	4.00	10.00	10.00	10.00	0	0	0	0	0	0
Intron 4	1857-1978	109	109	109	109				0	3	2			
Exon 5	1979-2035	57	16.17	16.17	16.17	40.83	40.83	40.83	0	0	0	0	0	1
Intron 5	2036-2145	102	102	102	102				0	2	0			
Exon 6	2146-2973	813	179.25	182.99	181.74	633.75	642.01	643.26	3	3	5	10	20	24
Intron 6	2974-3089	116	116	116	116				3	0	7			
Exon 7	3090-3533	426	88.89	89.21	89.45	334.11	333.79	333.55	1	2	4	2	3	8

Intron 7	3534-3859	255	255	255	255				5	14	0			
Exon 8	3860-3928	69	15.78	16.21	16.17	53.22	52.79	52.83	0	0	0	2	4	0
Total														
Coding	1-3928	1860	411.97	415.46	413.88	1442.03	1450.54	1452.12	5	8	11	16	32	36
Non-														
coding	1-3928	1927	1927.01	1928.01	1925.01	0	0	0	19	42	36	0	0	0
TOTAL	1-3928	3787	2338.98	2343.47	2338.89	1442.03	1450.54	1452.12	24	50	47	16	32	36

CC: *C. canadensis*; CS: *C. suecica*; CH: hybrid

Table 7. Sequence divergence between three groups

Interspecific	K_{all}	K_{silent}	$K_{noncode}$	K_s	K_a
CC-CH	0.00533	0.00526	0.00562	0.00358	0.00546
CS-CH	0.01191	0.01326	0.01285	0.01463	0.01004
CC-CS	0.01414	0.01549	0.01603	0.01216	0.01238

CC: *C. canadensis*; CS: *C. suecica*; CH: hybrid. K_{all}: the average number of nucleotide substitution per site between groups with Jukes-Cantor model for all sites; K_{silent}: the average number of nucleotide substitution per site between groups with Jukes-Cantor model for silent sites (synonymous and noncoding regions); $K_{noncode}$: the average number of nucleotide substitution per site between groups with Jukes-Cantor model for noncoding sites (flanking region and introns); K_s: the average number of nucleotide substitution per site between groups with Jukes-Cantor model for synonymous sites; K_a: the average number of nucleotide substitution per site between groups with Jukes-Cantor model for nonsynonymous site.

Table 8. Population recombination parameter (*4Nc*) estimates

Group name	# sequences (ω)	# subsamples	Gamma (γ)	Gamma (γ)/bp	*c/u*
'CC'	12	495	31.194	0.5045	1.5194
'CH'	26	2000	31.315	0.5277	1.9273
'CS'	7	35	32.629	0.5336	1.2894

CC: *C. canadensis*; CS: *C. suecica*; CH: hybrid. Gamma (γ) is the estimate of recombination rate, which equals to *4Nc*, where *N* is the effective population size and *c* is the recombination rate per generation. The number of subsamples of sequences is either equal to $\binom{\omega}{4}$ or 2000 random subsamples, whichever is fewer. *c/u* is the estimate of the number of recombination events per mutation event, which is equal to $4Nc/4N\mu$.

Table 9. ANOVA table of GLM test for all variables

Source	Degree of freedom	Sum of squares	Mean square	F value	P value
Model	14	49542.53	3538.75	10.56	<0.0001 ***
Error	32	10720.71	335.02		
Corrected total	46	60263.23			

* Significant at the $P<0.1$ level; ** significant at the $P<0.05$ level; ***significant at the $P<0.01$ level.

Table 10. GLM test of association between functional domain and petal color

Domain	# of variable sites included	Degree of freedom	Sum of square	Mean square	F value	P value
Interaction	5	4	607.25	151.81	0.44	0.7767
Acidic	4	3	4183.23	1394.41	4.07	0.0140**
bHLH	3	2	1940.90	970.45	2.83	0.0725*
C-terminal	3	2	825.19	412.59	1.20	0.3121

* Significant at the $P<0.1$ level; ** significant at the $P<0.05$ level; ***significant at the $P<0.01$ level.

Table 11. GLM test of the association between single amino acid site variation and petal color

Position/amino acid variation	Domain	sum of square	Mean square	F value	P value
11/E-G	Interaction	132.85	132.85	0.40	0.533
19/R-Q	Interaction	0.00	0.00	--	--
88/A-V	Interaction	6.49	6.49	0.02	0.890
94/Q-P	Interaction	127.21	127.21	0.38	0.542
98/R-K	Interaction	72.63	72.63	0.22	0.645
228/A-V	Acidic	0.00	0.00	--	--
249/D-V	Acidic	553.86	553.86	0.65	0.208
307/A-S	Acidic	3066.87	3066.87	9.15	0.005***, †<0.001***
380/D-G	Acidic	0	0	--	--
385/R-T	Acidic	119.49	119.49	0.36	0.555
436/S-L	bHLH	72.63	72.63	0.22	0.645, †0.0004***
464/V-I	bHLH	1267.18	1267.18	3.78	0.061*, †0.0073***
518/M-L	Linking region	0.00	0.00	--	--
549/V-I	C-terminal	27.59	27.59	0.08	0.776
623/W-C	C-terminal	87.03	87.03	0.26	0.613

* Significant at the $P<0.1$ level; ** significant at the $P<0.05$ level; ***significant at the $P<0.01$ level. †P values of GLM test are obtained using only variables from acidic domain and bHLH domain that are significantly associated with petal color.

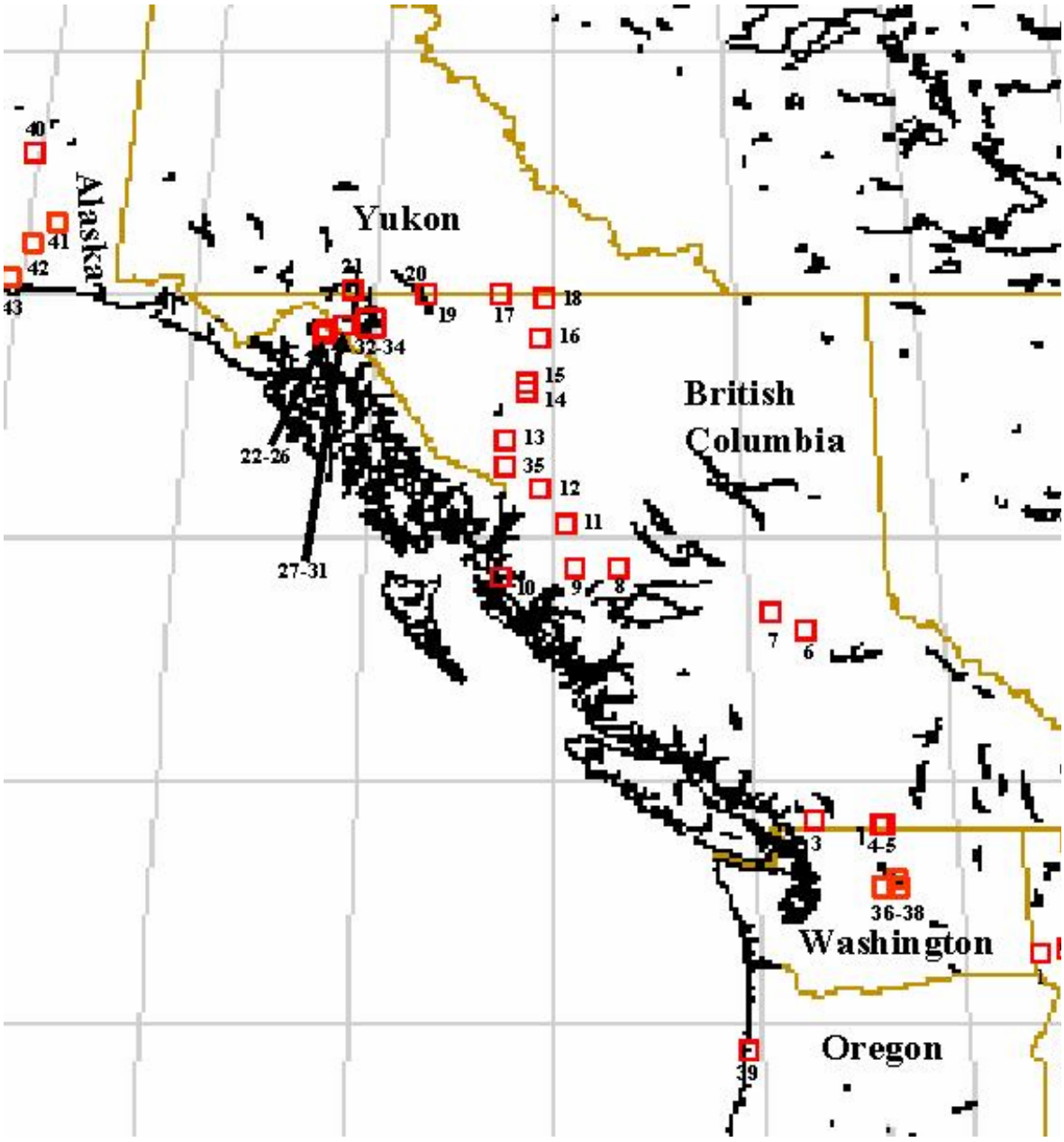


Fig. 1. Localities of 43 populations sampled.

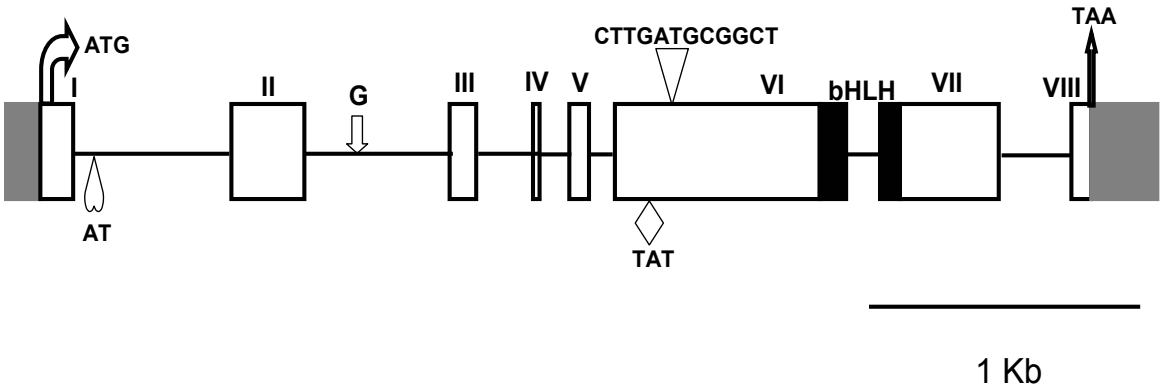


Fig. 2. Schematic map showing the overall structure of the anthocyanin regulatory gene for *C. canadensis* as deduced from full length nucleotide sequences. The position of the ATG translation start and the TAA translation stop codons are indicated. The region encoding bHLH is showing in black. The flanking regions are shown in shadow. Four indels are indicated by heart (2 bp), arrow (1bp), diamond (3 bp), and triangle (12 bp), respectively. The box represents exons, and the line represents introns. Exons are ordered as I-VIII.

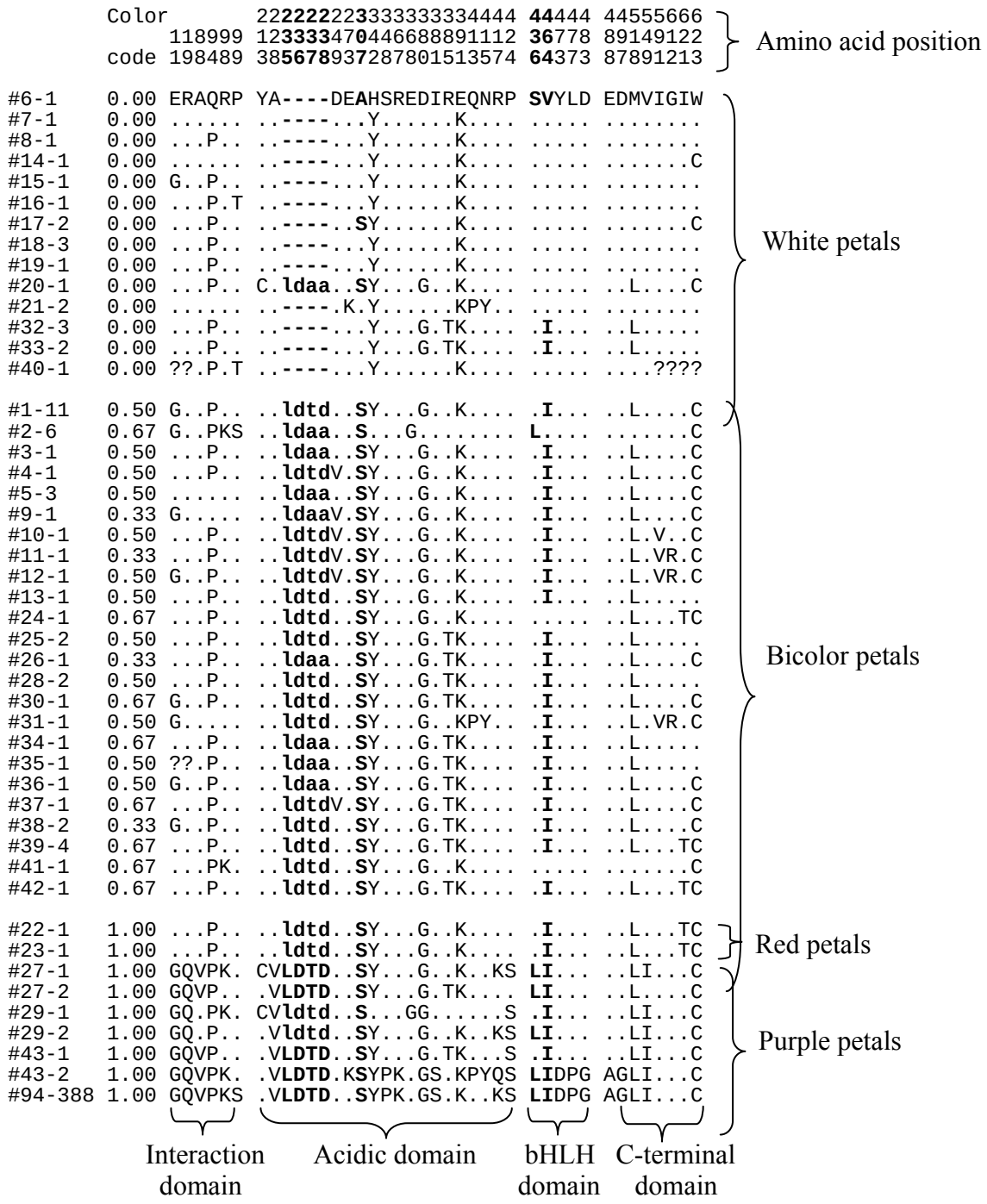


Fig. 3. Data matrix of score of petal color and informative amino acid sites of the *myc*-like anthocyanin regulatory gene among 47 samples of the dwarf dogwoods. Dots indicate identity to topmost sequence. Dashes represent gaps. Small-case letters indicates the

heterozygote of indel. The sites associated with petal color are bolded. Interaction domain: Site 1-193; acidic domain: site 194-430; bHLH domain: site 431-486; C-terminal domain: site 520-626.

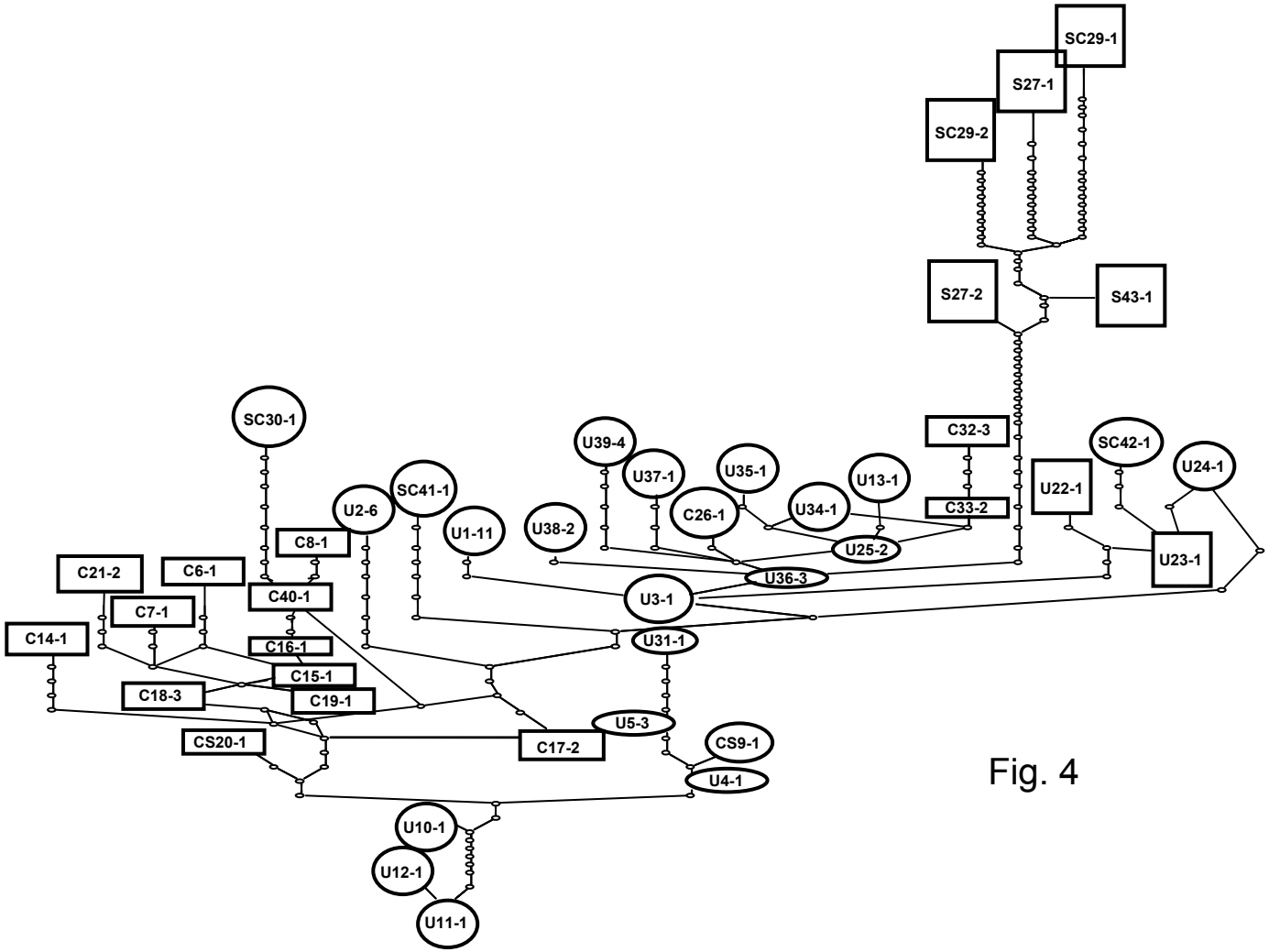

SC29-1
S27-1
SC29-2
S27-2
S43-1
C32-3
SC42-1
U24-1
U22-1
C33-2
U23-1
SC30-1
U39-4
U37-1
U35-1
U13-1
SC41-1
U2-6
C26-1
U34-1
U1-11
U38-2
U25-2
C8-1
C21-2
C6-1
U36-3
C40-1
U3-1
C7-1
C14-1
U31-1
C16-1
C15-1
C18-3
C19-1
U5-3
CS20-1
C17-2
CS9-1
U4-1
U10-1
U12-1
U11-1

Fig. 4

Fig. 4. Statistical parsimony haplotype network of the *myc*-like anthocyanin regulatory gene in the dwarf dogwoods. Rectangles represent haplotypes with white petals; squares represent haplotypes with purple petals. Ovals represent haplotypes with bicolor petals. Small circles represent missing haplotypes with >95% statistical parsimony support. C: morphologically putative species of *C. canadensis*; CS: morphologically putative hybrids *C. canadensis*>C. *suecica* (lineage 'B'); U: morphologically putative species of *C. unalaschkensis*; SC: morphologically putative hybrids *C. canadensis*<*C. suecica* (lineage 'D'); S: morphologically putative species of *C. suecica*. The numbers are the collection identity (see Table 1).