ABSTRACT

ZHANG, WENHENG. Molecular evolution of floral homeotic B-class genes in the dogwood genus *Cornus* (Cornaceae) – gene duplication, selection, and coevolution. (Under the direction of Dr. Qiu-Yun Jenny Xiang).

Comparative study of floral homeotic gene evolution through reconstructing gene genealogy is recognized as an important step toward understanding the molecular genetic basis of morphological evolution. The floral homeotic B-class genes, *APETALA3* (*AP3*) and *PISTILLATA* (*PI*) encode MADS domain-containing transcription factors required to specify petal and stamen identities in *Arabidopsis*. My dissertation study investigates *AP3*- and *PI*-like gene evolution through genomic DNA in the dogwood genus *Cornus* (Cornaceae). Our studies show that ancient gene duplications occurred in *CorPI* genes but were lacking in the *CorAP3* gene lineage during diversification of *Cornus*. Reconstruction of the *PI* genealogy in *Cornus* based on partial genomic DNA sequences reveals a dynamic history of gene duplication and loss at different phylogenetic levels. Ancient duplication events produced two ancient paralogs, named *CorPI-A* and *CorPI-B*, followed by the subsequent loss of gene copies in different subgroups during the early radiation of the genus. The *AP3* genealogy of *Cornus* found no ancient gene duplication during diversification of the genus, while two divergent gene copies are found in *Davidia*, a close relative of *Cornus*. Multiple sequence types are observed for both *CorPI* and *CorAP3* genes within the species of *Cornus*, suggesting that frequent, independent, and recent gene duplications occurred within species. Functional constraints of B-class genes in *Cornus* are generally relaxed compared to those observed in other plant groups. More relaxed functional constraints are found in *CorPI* than in *CorAP3*, and in *AP3*- than in *PI*-like genes in the outgroup genera *Alangium* and *Davidia*. Difference in the strength of selection on these paralog loci may be related to maintenance of

ancient paralog gene copies in the *CorPI* gene lineage in *Cornus*, and also in the *AP3*-like gene lineage in *Davidia*. The four positively selected amino acid sites detected on the *CorPI* genes indicate that the relaxation of functional constraints in *CorPI* may be due to diversifying or positive selection. Moreover, positive selection acting on these sites may be related to gene duplication. For the *CorAP3* loci, however, the relaxed functional constraint may be due to neutral evolution. I also discuss the divergence patterns of these two nuclear genes in *Cornus* and their potential use as phylogenetic markers.

Molecular evolution of floral homeotic B-class genes in the dogwood genus *Cornus* (Cornaceae) – gene duplication, selection, and coevolution

by

Wenheng Zhang

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

Botany

Raleigh, NC

November, 2006

Approved by:

Dr. (Jenny) Qiu-Yun Xiang Chair of Advisory Committee	Dr. Michael D. Purugganan Co-chair of Advisory Committee
Dr. Nina S. Allen	Dr. Brian M. Wiegmann

DEDICATION

To my parents

BIOGRAPHY

Wenheng became interested in plant systematics and later evolutionary biology in her third year of college. She became fascinated with evolution because it encompasses so many disciplines, such as geology, biology, and philosophy. Her master's thesis, focusing on these interests, was a molecular systematic study of the Dipsacales, a major lineage in the Asterids, the largest group of flowering plants, under the direction of Dr. Zhiduan Chen. The work was published in *Molecular Phylogenetics and Evolution*. After receiving her master's degree, Wenheng joined Dr. Jenny Xiang's plant systematics lab in 2002. Wenheng's studies have encompassed broad evolutionary topics, such as species level phylogeny of *Cornus*, comparison of the rates of speciation and molecular evolution in plants of eastern Asia and eastern North America, and biogeographical analysis and molecular dating of cornelian cherries. As a co-author, she published three papers based on these studies. Her dissertation topic is a study of the molecular evolution of the floral homeotic B-class gene in *Cornus*. Wenheng also enjoys teaching. As a teaching assistant in the Biotechnology program at NCSU, she won the 2006 University Outstanding Teaching Assistant Award. Born in Beijing, China, Wenheng earned her bachelor's degree in pharmacy, with honors, from Beijing Medical University in 1998 and her master's degree from Beijing University in 2001. Wenheng's long-term goal is to teach, continue her research and to work closely with students and colleagues to find answers to questions on plant systematics, molecular evolution, and evolutionary developmental biology. She married Jun He in 2002. Jun is now a graduate student in the Department of Bioinformatics and Computational Biology at George Mason University.

ACKNOWLEDGMENTS

My deepest gratitude goes to my advisor, Dr. Jenny Xiang, for her insight, guidance, enthusiasm, support, and patience. Although Jenny became sick before my defense, she continued to help with my dissertation revision. I am grateful to my committee members, Dr. Michael D. Purugganan, Dr. Brian M. Wiegmann, and Dr. Nina S. Allen for their insightful comments on my research and for their encouragement. I especially want to thank Dr. Brian M. Wiegmann who extended considerable effort toward improving the English in my dissertation. Many thanks also go to the current and former members of the Xiang Lab, David Thomas, Chuanzhu Fan, Kemei Ding, Jennifer Modliszewski, Melinda Peters, Alexandra Krings, Kathy McKeown, Chunmiao Feng, and AJ Harris for their friendship and for sharing their expertise. I particularly want to thank the following people who gave me help with experiments and data analyses: David Thomas for assistance with sequencing; Kemei Ding for help with collecting some of the *AP3* data, and Jun He for helping with histogram analysis and generating figures using the Excel program. I thank Dr. Jer-Ming Hu and Michael Frolich who provided primers and mRNA sequence data of *Cornus florida* and *Cornus alba* for this study. My special thank you goes to Dr. David E. Boufford for helpful discussions and for kindly proofreading the dissertation. I also want to thank the Deep Time Program, the Department of Botany and University Graduate Student Association for travel support during my study. This study was supported in part by National Science Foundation grants to Q-Y (J.) Xiang (NSF-DEB 0129069 and DEB 0444125).

TABLE OF CONTENTS

LIST OF TABLES

I. Molecular evolution of *PISTILLATA*-like genes in the dogwood genus *Cornus* (Cornaceae)

LIST OF FIGURES

I. Molecular evolution of *PISTILLATA*-like genes in the dogwood genus *Cornus* (Cornaceae)

LIST OF ABBREVIATIONS

aa - amino acid
AP3 - *APETALA3*
BB - big-bracted dogwoods
bp - base pair
BW - blue- or white- fruited dogwoods
CC - cornelian cherry
CorPI - *PISTILLATA*-like genes in *Cornus*
CorPI-A - A copy of *CorPI* in *Cornus*
CorPI-B - B copy of *CorPI* in *Cornus*
dN - nonsynonymous substitution
DNA - deoxyribonucleic acid
dS - synonymous substitution
dNTP - dinucleotide triphosphate
DW - dwarf dogwoods
GTR - general time reversible
kb - kilobase
LRT - likelihood-ratio test
MEGA - molecular evolutionary genetic analysis
ML - maximum likelihood
MP - maximum parsimony
MYA - million years ago
NJ - Neighbor-Joining method
PAML - phylogenetic analysis by maximum likelihood
PAUP - phylogenetic analysis using parsimony
PCR - polymerase chain reaction
PEG - polyethylene glycol
PI - *PISTILLATA*
RFLPs - restriction fragment length polymorphisms
SD - standard deviation
TBR - tree bisection reconnection

Chapter I

Molecular evolution of *PISTILLATA*-like genes in the dogwood genus *Cornus* (Cornaceae)

Keywords: adaptive evolution, *Cornus*, gene duplication, MADS-box genes, molecular evolution, phylogeny, *PISTILLATA*-like genes, regulatory gene evolution.

Abbreviations

PI: *PISTILLATA*

CorPI: *PISTILLATA*-like genes in *Cornus*

CorPI-A, *CorPI*-B: two major copies of *CorPI* in *Cornus*

Abstract

The *PISTILLATA* (*PI*) gene is a member of the plant MADS-box gene family responsible for specifying differentiation of petals and stamens during the development of angiosperm flowers. The MADS-box gene family encodes critical regulators determining floral organ development. Understanding evolutionary patterns and processes of the MADS-box genes is an important step toward unraveling the molecular basis of floral morphological evolution. However, the evolutionary history, pattern, and processes of MADS-box genes in most flowering plants remain unknown. In this study, we investigated *PI*-like gene evolution in the dogwood genus (*Cornus*, Cornaceae), a eudicot lineage in the clade Asteridae. Our reconstruction of the genealogy, based on genomic DNA sequences of one third of the 5' end of the *PI*-like gene, revealed a dynamic history of gene duplication and loss at different times in *Cornus*. We detected ancient duplication events, especially two ancient paralogs, named *CorPI*-A and *CorPI*-B, followed by subsequent losses of gene copies in different subgroups during the early radiation of the genus. Each species analyzed was found to contain multiple copies with most of them derived recently within species, suggesting frequent independent gene duplications in all species. These results are similar to those found in recent studies of the basal eudicots, e.g., Ranunculaceae. Estimation and comparison of dN/dS ratios using both MEGA and PAML programs revealed a relaxation of selection (measured by dN/dS ratio) in the *PI*-like gene in *Cornus*. Measured dN/dS ratios are significantly greater in *Cornus* than in closely related outgroup genera *Alangium* and *Davidia*, and than those reported for other MADS-box genes from other flowering plants, suggesting relaxed selective constraints in the *PI*-like paralogous genes in *Cornus*. Strong positive selection at several amino acid sites of *CorPI* were also detected, with most of these sites from the I-

domain of the gene, a region critical for dimerization activity. Changes at those positively selected sites involved replacements of amino acids differing in charges and polarities, which suggested possible functional changes of the genes. Selection on *CorPI* genes also differs among major copies and among the *Cornus* lineages. Functional constraints in the *CorPI*-A copy that was preserved in all lineages of *Cornus* are more relaxed than in the *CorPI*-B copy. The two *Cornus* lineages with petaloid bracts exhibit more relaxed selection in *PI*-like genes when compared to other lineages. The total substitution rates of the *PI*-like gene also differ among lineages, showing a trend similar to that found with the dN/dS ratio. Evolutionary rates and dN/dS ratios in these two lineages appear to correlate with the origin of petaloid bracts in the two lineages and with the single herbaceous habitat of the DW lineage. Finally, we found that the *CorPI*-A copy contains informative phylogenetic information when compared across *Cornus* species.

Introduction

The MADS-box gene family encodes a series of transcription factors involved in controlling the formation of flowers, flowering time, and vegetative development in plants (Jack 2001, Ng and Yanofsky 2001). The landmark ABC model of flower organ identity includes three classes of homeotic genes that mostly encode MADS-box proteins (Coen and Meyerowitz 1991; Schwarz-Sommer et al. 1990). These genes function in combination to specify regional identities in the four floral whorls of a flower (Schwarz-Sommer et al. 1990; Coen and Meyerowitz 1991; Weigel and Meyerowitz 1994). That is, the A class genes alone specify sepals, the A and B class genes together specify petals, while the B and C genes

together determine stamens and the C class gene alone determines carpels. Recent discoveries have extended the ABC model, i.e., the 'quartet model', to include D and E classes of genes which are found to be involved in development of the ovary and floral organs (Pelaz et al. 2000; Theissen and Saedler 2001). Studies of selected monocots (e .g. rice and maize) and eudicots (e.g. *Arabidopsis* and *Antirrhinum*) show that the ABC model appears to be generally conserved across angiosperms (Ambrose et al. 2000; Whipple et al. 2004). In contrast, studies also demonstrate functional divergence of the ABC program across angiosperms (Kramer and Irish 1999, 2000; Theissen et al. 2000), suggesting that the evolution of floral identity genes and modification of the flower programs may be the key factor to flower evolution. Therefore, study the evolution of the MADS-box genes and the ABC flower program provides an opportunity to understand the genetic basis of floral evolution during the diversification of angiosperms.

Recent studies reveal that MADS box genes have experienced extensive gene duplications (Parenicova et al. 2003; Nam et al. 2004; Irish and Litt 2005). Key duplication events were shown to correlate with major radiations of the seed plants (Purugganan 1998; Becker et al. 2000; Theissen et al. 2000). The gene duplications followed by functional divergence of the paralogs in the MADS-box gene lineage were considered to have resulted in modifications of the ABC program (Irish and Litt 2005). Functional divergence of paralogous MADS-box genes has also been implicated as a possible major factor in adaptive evolution of floral morphology in angiosperms (Lamb and Irish 2003; Litt and Irish. 2003; Kramer et al. 2004; Aoki et al. 2004; Stellari et al., 2004; Vandenbussche et al. 2004; Di Stilio et al. 2005; Kim et al. 2005a; Kim et al. 2005b; Zahn et al. 2005).

Molecular evolutionary studies have explored the relationship between sequence evolution of MADS-box genes, phenotypic variation, and species diversification in plants (Barrier et al. 2001; Martinez-Castilla and Alvarez-Buylla 2003). A recent study detected adaptive sequence evolution of the MADS-box gene family in *Arabidopsis*, and the evidence suggested that those sites under positive selection may have played important roles during MADS-box gene diversification through acquisition of novel functions, as well as during phenotypic evolution of plants (Martinez-Castilla and Alvarez-Buylla 2003). Other studies have also suggested that evolution of regulatory genes in general must play a key role in the morphological divergence of plants (Lowe and Wray 1997; Doebley and Lukens 1998; Purugganan 1998). Major lines of evidence leading to these hypotheses include elevated ratios of replacement to non-replacement substitution rates (ω=dN/dS) in regulatory genes, compared to structural genes, and diversifying or directional selection detected in regulatory gene coding regions (Purugganan 1998; Barrier et al. 2001; Remington and Purugganan 2002). Furthermore, molecular studies of development show that dramatic shifts in organismal structure can arise from mutations at key regulatory loci (Lowe and Wray 1997).

Positive selection in the MADS-box gene family (in both MADS- DNA-binding and K-non-DNA-binding domains) of *Arapidopsis* and in the HOX gene non-DNA-binding region of vertebrates were both found to be associated with gene duplications (Martinez-Castilla and Alvarez-Buylla 2003; Van de peer et al. 2001; Fares et al. 2003). Alternatively, the evolution in the non-DNA binding region of a single ortholog of an anthocyanin regulatory gene (a structural gene activator, *Ipmyb1*) was found to be neutral in *Ipomea* (Chang et al. 2005).

Therefore, it has been hypothesized that positive selection may have contributed to the rapid evolution of regulatory genes immediately following gene duplication, while subsequent rapid evolution within a single ortholog regulatory gene may be primarily neutral (Chang et al. 2005). Clearly, data from additional regulatory genes in other organisms will be helpful to provide a better understanding of gene duplication associated sequence evolution.

PISTILLATA (*PI*) is classified as a B-class gene of the MADS-box gene family. *PI* homologs function together with another B-class gene, *APETALA3* (*AP3*), by forming heterodimers in the pathway of regulating petal and stamen development in eudicot plants (Schwarz-Sommer et al. 1992; Tröbner et al. 1992; Goto and Meyerowitz 1994; Riechmann et al. 1996a; Riechmann et al. 1996b). B-class gene functions appear to be conserved across the orthologs analyzed among the core eudicots (Jack et al. 1992; Goto and Meyerowitz 1994) and monocots (Ambrose et al. 2000; Whipple et al. 2004). Like most plant MADS-box genes, the *PI* homologs consist of four conserved modules, known as the MIKC structure (Alvarez-Buylla et al. 2000, Fig 1). The core region of the plant MADS-box protein has been defined to consist of the MADS-box, the I-region, and the first part of the K-region. These regions are necessary for DNA-binding and dimerization activities (Krizek and Meyerowitz 1996a, Krizek and Meyerowitz 1996b; Riechmann et al. 1996a; Riechmann et al. 1996b; see Figure 1). The other half of the protein, the noncore region, includes the rest of the K-box and the C-terminal sequence involved in strengthening specific dimerization activities (Riechmann and Meyerowitz 1997, Lamb and Irish 2003).

Several evolutionary studies of B-class genes have been conducted recently, largely on basal angiosperms and lower eudicot lineages that have highly diverse floral morphology (Kramer et al. 1998; Kramer et al. 2003; Stellari et al. 2004; Kim et al. 2004; Aoki et al. 2004). These studies reveal that the *PI* homologs exist as a single copy gene in most of the species investigated. However, two or more copies of the gene were also found in some species (Kramer et al. 1998; Kramer et al. 2003; Stellari et al. 2004). Evidence available thus far indicates that *PI*-like gene duplications occurred at various phylogenetic levels, within species, within a genus, and in ancient genes found in the ancestors of clades of families and orders, suggesting a dynamic pattern of *PI*-like gene evolution (Kramer et al. 1998; Kramer et al. 2003; Stellari et al. 2004). At present, it is unknown if the evolutionary pattern of B-class genes observed in these basal angiosperms and lower eudicots represents a common phenomenon in all angiosperms. It is, therefore, necessary to investigate other angiosperm lineages to determine if the pattern of *PI* gene duplication observed in lower eudicots also occurs in higher eudicots and in lineages without modifications of floral organs (e.g., petaloid sepals, lacking petals or sepals etc. as found in studies of the lower eudicot, Ranunculales). Furthermore, previous studies of *PI* gene evolution were conducted only at the mRNA sequence level and no examination of genomic DNA evolution has been conducted on MADS-box genes. Clearly, mRNA sequences alone can only reveal part of the story of a gene's history. Analyses of genomic DNA sequences provide a broader context within which to investigate the evolutionary history of a gene or gene family.

In the present study, we report our findings on genomic sequence evolution of the *PI* gene in the dogwood genus *Cornus* (Cornaceae), a lineage of asterid eudicots in the

Angiosperm tree of life. Cornaceae radiated into four morphologically distinct subgroups in the early Tertiary (at least 50-60 million years ago based on fossil data) that have diverged in their inflorescence architecture and bract morphology (Eyde 1988; Xiang et al. 2006). Flowers of all species of dogwoods produce normal flowers (with sepals, petals, stamens, and pistil) except in one species in which the stamens and pistils become sterile on different plants (a reproductive system of functional dioecy). Two of the four subgroups (e.g., the flowering dogwood, *Cornus florida*, and the bunch berry, *C. canadensis,* and their allies) evolved four petaloid bracts at the base of the inflorescence. The bracts in the other subgroups are not petaloid, but leafy or scale-like. Showy bracts are key trait in flowering dogwoods (*C. florida*, *C. kousa*), contributing to their enormous horticultural value.

Dogwoods are an excellent system for investigating floral gene evolution for several reasons: 1) the genus has a well established phylogeny based on previous studies of multiple genes (Xiang et al. 1993, 1996, 1998, 2002, 2006; Fan and Xiang 2001, 2003), thus providing an important framework for elucidating gene evolution; 2) given the presence and absence of petaloid bracts in the genus, species-level investigation permits exploration of the possible relationship between the evolution of the gene and the evolution of petaloid bracts; **3)** the genus represents a lineage of higher eudicots that does not exhibit abnormality in petals and stamens among the species, thus providing evidence of *PI*-like gene evolution in taxa different from the basal angiosperms and lower eudicots; and 4) the subgroups of the genus are old enough to allow observation of sequence divergence of coding regions, but likely not too old to erase substitutions at synonymous sites due to mutational saturation. Through analysis of genomic DNA sequences with multiple sampling from each subgroup and its

constituent species, we provide a detailed genealogy of the *PI*-like genes of *Cornus* to study the following aspects of gene evolution: (1) dynamics of gene duplication and loss; (2) levels and variation of selection among major lineages of *Cornus*, among paralogous gene copies, and among amino acid sites; and (3) rate heterogeneity of molecular evolution of *CorPI* among species lineages through time.

Materials and Methods

Taxon Sampling

The genus consists of ~55 species divided among four major lineages. The blue- or white- fruited group (BW) is the most diverse and produces open, compound cymes with minute, early-deciduous green bracts. The cornelian cherry group (CC) has 6 species, producing umbellate cymes with four basal, scale-like bracts. The big-bracted group (BB) contains ~8-13 species that possess capitate cymes subtended by four, large, petaloid bracts except in one species (*C. disciflora*) which lacks the petaloid bracts. The last group is the dwarf dogwoods (DW), the only herbaceous lineage in the genus, with 3 species that bear minute compound cymes subtended by four petaloid bracts. We analyzed 24 samples (Table 1) of 16 species of *Cornus* from the four subgroups (Fan and Xiang 2001; Xiang et al. 2006). We sampled multiple individuals of the same species and multiple species from the same subgroup to confirm and detect variation in the copy numbers of *CorPI* within species and within subgroups, and to distinguish sequence variants from the PCR recombination and intergenic recombination processes. Two genera, *Alangium* (Alangiaceae) and *Davidia* (Davidiaceae), close relatives of *Cornus* based on phylogenetic analyses of Cornales (Xiang

et al. 1998, 2002; Fan and Xiang 2003), were used as outgroups (Table 1). *Alangium* is the sister genus of *Cornus*, and *Davidia* is in a clade sister to the *Alangium-Cornus* (Xiang et al. 1998, 2002; Fan and Xiang 2003). *PI* homologs of other eudicots (*Arabidopsis thaliana* NM122031 and *Antirrhinum majus* X68831) and *PI*-like genes of lower eudicots and basal angiosperms (*Sanguinaria canadensis* AF130871; *Dicentra eximia* AF052857; *Piper magnificum* AF052866; *Peperomia hirta* AF052865; *Meliosma dilleniifolia* AY436712; *Calycanthus floridus* AF230708; *Thottea siliquosa* AY436708; *Thottea siliquosa* AY436717; *Houttuynia cordata* AY436746; *Houttuynia cordata* AY436707; *Lindera erythrocarpa* AY436739; *Chloranthus spicatus* AF230710; *Nymphaea sp.* AY436744), were downloaded from Genbank for comparison of *PI* homolog diversification between *Cornus* and other flowering plants.

CorPI **Cloning and Sequencing**

CorPI loci were amplified from total genomic DNA extracted from leaves using the Qiagen DNeasy extraction kit (Qiagen, Germany) or the modified CTAB miniprep method (Xiang et al. 1998). A fragment of approximately 1200 base pairs was amplified using *Cornus*-specific *PI*-like gene PCR primers, which are located on the two conserved regions of the MADS-box domain (exon 1) and the K domain (exon 4) (Figure 1). These primers (Forward primer: CPI-1L1 5'TGTTATCTTTGSTAGCTCTGGCAAGAT3' and Reverse primer: CPI-KR 5'GTGATATCTTCCCCCTTCAGGTG3') were designed based on the cDNA sequences of *PI*-like genes from *C. florida* (BB group) and *C. alba* (BW group) (provided by Drs. Jer-Ming Hu and Michael Frolich unpublished data). The PCR reaction contains 5μL of 10X Mg^{2+} free buffer (Promega), 6μL of 25mmol/L $MgCl_2$ (Promega), 10μL

of 2.5mmol/L of each dNTPs, 1μL of 20μmol/L forward primer, 1μL of 20μmol/L reverse primer, 5μL of BSA (Bovine serum albumin, 10mg/ml), 1.5 units of Taq polymerase (Promega), 5-10μL of 20ng/μL total DNA extract, and calibrated to a final volume of 50μL using sterile deionized water. A hot-start step, 6 min of 96 ^{o}C incubation, was used to denature the genomic DNA templates before adding Taq polymerase (Promega). PCR was then performed at 94 ^{o}C for 1 min followed by 36 cycles at 94 ^{o}C (30 s), 62 ^{o}C (1 min), 72 ^{o}C (2 min) and a final step of 7 min extension at 72 ^{o}C. After gel purification of the PCR products using Qiaquick gel extraction kit (Qiagene, USA), direct sequencing produced unreadable sequences suggesting that more than one sequence type of the gene exists in the samples.

To separate allelic and homologous versions of the gene, a cloning procedure was employed. The PCR products were first cleaned using 20% PEG (polyethylene glycol) 8000/2.5mol/L NaCl to increase ligation efficiency before applying the TOPO TA cloning kit with competent cells following the manufacturer's protocol (Invitrogen). Positive transformants were detected by the correct insert size through PCR screening using universal T3 forward and T7 reverse primers provided by the cloning kit. For most samples, 40-63 PCR products of positive transformants from each sample (Table 1) were processed through Restriction Fragment Length Polymorphisms (RFLPs) analysis using restriction endonucleases Taq I (Promega) and Hsp92 II (Promega) for preliminary detection of differing sequence types. For eight samples, 10-20 clones were analyzed by sequencing (Table 1). The low number of clones screened for these eight samples did not limit recovery of the real pattern of *CorPI* evolution, because additional samples from each species or

samples from closely related species were exhaustively screened (Table 1). The restriction endonucleases were chosen based on sequence data of 10 clones from *C. florida* (Voucher 02-16), which cleave differently among different copies of the *PI*-like genes. One to three positive clones representing each banding pattern in the RFLPs analysis were inoculated in a nutrient medium to multiply the cells. Plasmid DNAs with *PI*-like gene inserts were extracted and purified using the Minipreps DNA purification system (Promega) for subsequent sequencing analyses (PerkinElmer). Forward and reverse sequencing was performed to control error in Taq polymerase (PerkinElmer) replication of the target sequence using T3 forward and T7 reverse primers (or CPI-1L1 forward and CPI-KR reverse primers). Two hundred and twenty five clones were sequenced and all sequences obtained were deposited in Genbank.

Phylogenetic analyses

Clone sequences from each sample were first analyzed using the Bellerophon program (Huber et al 2004) to detect chimeric sequences. Five sequences were found to be chimeric, which may have formed through the PCR recombination process and these were excluded from subsequent analyses.

An initial phylogenetic analysis was performed including all *PI*-like genes from *Cornus* and other flowering plants (see Materials and Methods, taxon sampling) to test the homology of *Cornus PI*-like genes (*CorPI*) on a broader scale. Exons alone were used in this analysis because introns could not be reliably aligned between sequences from *Cornus* and other families. One hundred and fifty sequences representing all sequence variants according to

both coding and noncoding regions were included in this analysis. Phylogenetic analyses of this matrix were conducted using both Neighbor-Joining (NJ) and Bayesian Metropolis-Hastings coupled Markov chain Monte-Carlo (MCMC) methods to reconstruct the gene genealogy. ModelTest 3.06 (Posada and Crandall, 1998) was first used to select a best-fitting model of sequence evolution. The GTR + Γ model was recommended for this dataset, which assumes general time reversibility (GTR) and a gamma approximation of the rate-variation among sites (Γ). The parameters of the GTR + Γ model based on Modeltest 3.06 results were applied to subsequent NJ and Bayesian analyses. Neighbor-Joining analysis (Saitou and Nei 1987) using the Jukes-Cantor model was performed using PAUP* 4.0b10 (Swofford 2002) with 10,000 bootstraps. For Bayesian likelihood analysis, the MCMC phylogenetic analyses were carried out using MrBayes 3.1.2 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) with flat priors and three heated chains in addition to a single cold chain. MCMC analyses were conducted in two parallel runs simultaneously initiated with a random tree, and run for a total of 1,000,000 generations (sampling trees every 100 generations) until a good sample from the posterior probability distribution was reached. Both searches reached a stationary state after about 110,000 generations which were detected using Tracer v1.1 (Rambaut and Drummond 2004). Conservatively, the first 1,100 trees from each run were discarded as burn-in and were not included in generating the consensus phylogeny. Summary statistics and consensus phylograms with nodal posterior probability support were estimated from the combination of trees after burn-in from the two runs per analysis. In both neighbor-Joining and Bayesian analyses, the trees were rooted using lower eudicots and basal angiosperms (Fig 2).

Since two major lineages of *PI* paralogs, *CorPI*-A and *CorPI*-B were identified in *Cornus* based on the broad phylogenetic analyses of exon sequences, and the intron sequences between the two sequence types cannot be aligned, detailed phylogenetic analysis is reported here only for *CorPI*-A (retained in all four subgroups and with multiple copies within a species) including both intron and exon sequences. *CorPI*-B was detected in only two of the subgroups and has few copies (Fig 2, 3). *CorPI*-A sequences were aligned using Clustal X (Thompson et al. 1997) and adjusted by eye. Modeltest indicated that GTR + Γ model is the best model to fit the data. Both NJ and Bayesian analyses were performed as described above without rooting. The Bayesian method was run in four chains with flat prior probabilities for 500,000 generations with sampling of every 100 generations. The first 1,000 trees (20% of the total) were discarded as the burn-in phase.

Estimating the strength of selection and detecting positive selection in the *PI*-like genes of *Cornus*

A total of 71 distinct coding sequences were detected in the study and are included in the selection analyses. The sequences covered the complete I domain and partial MADS- and K- domains of the gene (Fig 5). First, we tested if selection (dN/dS ratio) acts differentially along branches associated with paralog gene copies or major morphological subgroups. Three questions are examined: (1) does *CorPI* gene have a different rate of evolution in *Cornus* than in its close relatives *Alangium* and *Davidia*?; (2) do the rates between the two paralog genes *CorPI*-A and *CorPI*-B of *Cornus* evolve differently after gene duplication; (3) whether the rates among the four different subgroups of *Cornus* evolved differently. First, we measured the pairwise dN/dS values using the Nei and Gojobori (1986)'s method under the

Jukes-Cantor model as implemented in MEGA 3.1 (Kumar et al. 2004) for all *CorPI* sequences, *PI*-like genes in outgroups (*Alangium* and *Davidia*), and orthologues *CorPI*-A and *CorPI*-B separately (Fig 6). We then examined the variation of dN/dS ratios of *CorPI*-A within each of the four major subgroups of *Cornus* (Fig 6). The statistical significance of differences in dN/dS values among lineages was examined by using the two-sample bootstrap resampling test with 10,000 replicates. We applied the Tajima relative rate test to detect for dS rate variation among the different subgroups in *Cornus* using *Davidia* and *Alangium* as outgroups (Tajima 1993) to determine if the differential dN/dS values observed are correlated with differential neutral evolutionary changes. The dN/dS values were then plotted against the dS values to detect the pattern of dN/dS ratio variation in relation to dS (Fig 7).

To further examine the level and variation of selection (dN/dS ratio) along different lineages (eg. paralogous genes or morphological subgroups), we measured the dN/dS values using the codon-based likelihood models (Table 2, models reference Yang 1998) in the Codeml program implemented in PAML 3.15 (Yang 1997). The Bayesian tree topology, which is also consistent with the NJ tree, was applied as a framework phylogeny for the analyses (Fig 8). The branch lengths estimated under the model, assuming one dN/dS ratio for all sites (Goldman and Yang 1994), were applied for the tree topology in later PAML analyses. We first applied the model that allows all 133 branches on the phylogeny to have different dN/dS ratios to estimate the dN and dS values for all branches to detect positive selection on the branches. To examine whether selection differs between *CorPI*-A and *CorPI*-B and between *Cornus* and its outgroups *Alangium* and *Davidia*, the dN/dS ratios

were estimated for under three different models, the three-ratio model, the simple model, and the two ratio model (Table 2; Fig 7) (method and models refer to Yang 1998). The three-ratio model assumes a different dN/dS ratio for each of the three defined branches, i.e., *CorPI*-A (ω_1), *CorPI*-B (ω_2), and outgroups (ω_0) (Table 2, A). The simplest model assumes one dN/dS ratio among the three branches ($\omega_1 = \omega_2 = \omega_0$, Table 2, B). Three two-ratio models (Table 2, C-E) assume that one of the branches has different ratio. For example, model D assumes that branch *CorPI*-A has a different ratio (ω_1) while all other branches have the same background ratio ($\omega_2 = \omega_0$). These models were then compared using the likelihood ratio test (LRT) by twice comparing the log-likelihood difference between two models to a chi2 distribution, with the number of degrees of freedom equal to the difference in the number of free parameters between the two models to examine hypotheses regarding existence of differential selection force between outgroup lineages and *CorPI* genes in *Cornus*, and between *CorPI*-A and *CorPI*-B (method reference to Yang 1998, Table 3). This codon-based likelihood approach was also applied to examine differences in selection on *CorPI*-A among the four morphological subgroups of *Cornus* and for comparison with the results obtained from MEGA using dN/dS on pairwise comparisons. .

The pairwise comparison approach and the codon-based likelihood method applied above estimates substitution rates to detect positive selection by averaging over all sites of the protein. However, adaptive evolution probably affects only a few amino acid sites due to the structural and functional constraints acting on the protein (Yang et al. 2000; Yang and Nielsen 2002), therefore averaging rates over sites may not be powerful enough to detect positive selection. We also examined dN/dS values at individual amino acid sites to

determine if adaptive molecular evolution (or positive selection) occurs at specific sites. Coding sequences of 67 *ConusPI* genes were applied in this analysis. Six codon-substitution models were used as implemented in the Codeml program of PAML 3.15 (Yang 1997, Yang et al. 2000, Table 4). We compared the models M0 (one-ratio model, ω values are the same at all amino acid sites), M1 (neutral model, all amino acid sites are either completely constrained ω = 0 or neutral ω = 1), and M7 (beta model, ω values are limited to the interval between 0 and 1, and eight categories of sites with independent ω distributed according to a ß distribution), all three representing the null hypotheses of no positively selected sites allowed, with models M2 (selection model, similar to M1 with an additional category of amino acid sites under positive selection ω > 1), M3 (discrete model, three categories of unconstrained ω of amino acid sites were estimated from the data), and M8 (beta & ω model, adds an extra site category to M7 with a free ω ratio estimated from the data allowing ω >1) representing the alternative hypotheses that allow for heterogeneous and positive selection (Yang et al. 2000). By comparing these nested models using the LRT (i.e., M0 vs. M3, M1 vs. M2, and M7 vs. M8), we could determine whether models allowing for positive selection fit the data better, indicating positive selection on the gene, and identify which amino acid sites are under positive selection (Yang et al. 2000). All analyses for detecting positively selected amino acid sites were performed using Bayesian empirical analysis (Yang et al. 2005) implemented in the Codeml program in PAML (Yang 1997).

Because initial phylogenetic analyses suggested two alternative placements of *PI*-like genes relative to the outgroup taxon, *Davidia*, one being sister to *CorPI*-A and one being sister to *CorPI*-A-*CorPI*-B-*Alangium* (Fig. 2-A and -B), topologies depicting both

placements were applied in tests for heterogeneity of selection along branches and in tests fordiversifying selection at amino acid sites as described above using a codon-based likelihood method. Results from both analyses were similar, and only the results based on *CorPI*-A-*CorPI*-B-*Alangium-Davidia* relationships are reported. Previous study has shown that the codon-based substitution models for inferring positively selected amino acid sites are not sensitive to an assumed tree topology and have little effect on the power and accuracy of the analysis (Yang et al. 2000); a finding which is also supported by our analyses of the *CorPI* genes.

Rate of molecular evolution and divergence time estimation

We employed the penalized likelihood (PL) method in r8s 1.71 (Sanderson 2002) to estimate divergence times and absolute rates of substitution for full length *CorPI*-A, including both coding and noncoding regions at all nodes of the gene genealogy. Branch lengths were generated using the maximum likelihood (ML) method in PAUP* 4.0b10 (Swofford 2002) using a constraint tree topology concordant with both the NJ and Bayesian analyses. Fossil information of *Cornus* was used to place calibration points on the genealogy. The node between CC and BB-DW was fixed as 62 mya following Xiang et al. (2005). We also performed cross-validation analysis in r8s 1.71 to choose the optimal value of the rate smoothing parameter for our data (Sanderson 2002). If the smoothing parameter is large, the nucleotide substitution rate is reasonably clocklike; if it is relatively small, then greater rate variation is allowed. The estimation of absolute divergence time and of nucleotide substitution rate was performed using the penalized likelihood method with the TN algorithm suggested by the program (Sanderson 2002).

Results

Chimeric clones and Taq polymerase error

PCR recombination and Taq polymerase errors are recognized as intrinsic artifacts of the PCR based cloning approach during the separation of different sequence types. We identified seven clones [*C. florida* 01-1 (clone 15, 38, 61), *C. florida* 01-148 (clone 7), *C. canadensis* 04-01 (clone 1), *C. alba* (clone 15), *C. oblonga* 02-223 (clone 12)] to be chimeric sequences that may be caused by PCR artifacts detected by the Bellerophon program (Huber et al 2004). These sequences combine the first part of one sequence type with the second part of the other sequence type from the same sample. However, chimeric clones of *C. florida* 01-1 (clone 15) and *C. florida* 01-148 (clone 7) from two independent samples show the same split point in their chimeric sequences, suggesting that these sequences might be caused by intergenic recombination instead of PCR-mediated recombination. Except for these two clones, all other chimeric sequences detected were excluded from subsequent analyses.

Taq error in PCR occurs at a rate (mutation frequency/bp/duplication) of 8×10^{-6} (Cline et al. 1996). For a 1200 bp sequence in this study, there should be fewer than 1 (actual value 0.58) error expected. Thus, detected sequence differences among clones can be inferred to have arisen from alleles, orthologs, or paralogs. Furthermore, within species from the BB, BW, and DW subgroups we detected multiple sequence types having nucleotide substitutions in both the coding and intron regions, whereas few sequence types were detected from species from the CC subgroup. Differing degrees of polymorphism observed in different

species under the same experimental conditions further supports the inference that sequence polymorphisms found in this study were not caused by Taq error.

Structural characteristics and sequence variation

The novel sequences obtained in the study were confirmed to be *PI*-like genes using BlastN search (Altschul et al. 1997) against sequences in Genbank. The *Cornus PI*-like genes we sequenced range from 1.1 to1.3 kb among the sixteen sampled species, representing one third of the gene from the 5'end. This region spans the exons of partial MADS-box, complete I- and partial K- domains and the introns between them (Fig 1). The exon regions are conserved in size among species. The coding regions cover 204 bp, including 37 bp out of the 171 bp of the MADS domain, the complete 93 bp of the I-domain, and 77 bp out of the198 bp of the K-domain. However, intron regions vary greatly in length among species and among gene copies, ranging from 151 to 364 bp for intron 1, 555 to 715 bp for intron 2 and 54 to181 bp for intron 3.

The sequence variants of *PI*-like genes obtained in *Cornus* can be sorted into two major types (named *CorPI*-A and *CorPI*-B) based on the gene genealogy (Fig 2 and 3; see below). The two major copies can also be easily distinguished by their high degree of divergence in the intron sequences they contain, through nucleotide substitutions and length differences, eg. in *C. chinensis*, *C. mas*, *C. officinalis*, *C. suecica* and *C. unalaschkensis* (Fig. 4). The high variability of their introns makes alignment of these regions complex, or even impossible, to infer.. In contrast, all of the exon sequences are easily aligned between the two copies and between *Cornus* and the outgroup taxa. The exons from the species of *Cornus* alone contain

30.39% (62/204) parsimony informative sites and 50.49% (103/204) variable sites. When *Alangium* and *Davidia,* or other eudicots and basal angiosperms as well as *Alangium* and *Davidia,* are included (see Materials and Methods, Taxon sampling), these numbers increase to 53.43% (109/204) and 86.27% (176/204) for variable sites and 35.2% (72/204) and 65.20% (133/204) for parsimony informative sites, respectively. Matrices including both exon and intron sequences were constructed for the *CorPI*-A copy, which is retained in all major lineages of *Cornus*, and for the *CorPI*-B copy, which failed to be detected in two major lineages of *Cornus* (see below). The *CorPI*-A matrix contains 1835 bp aligned sequences, of which 42.02% (771/1835) are parsimony-informative, and 55.42% (1017/1835) are variable. The final alignment of *CorPI*-B is 1100 bp in length, with 37.5% (413/1100) of sites parsimony-informative and 43.3% (476/1100) of sites variable.

Genealogy and gene duplication

Phylogenetic analysis of coding sequences revealed that the *CorPI* genes form two monophyletic clades, each including representatives of different subgroups of *Cornus* (Fig. 2). This further indicates the presence of two major copies (*CorPI*-A and *CorPI*-B) in *Cornus* that arose from an ancient gene duplication event at least before the initial diversification of *Cornus* (Fig. 2). In NJ analyses, the *CorPI*-A and *CorPI*-B lineages are grouped as sister to the *PI* homologs from *Alangium* and *Davidia,* outside of the genus *Cornus* (Fig. 2-A). The ((*Cornus*, *Alangium*), *Davidia*) relationship was also supported by phylogenetic analyses based on other molecular markers (Xiang et al. 2006). In Bayesian analyses, the *PI*-homologs from *Davidia*, a monotypic genus that also has large petaloid bracts, is allied with the *CorPI*-A clade (Fig. 2-B). However, neither of these placements of *Davidia* is strongly supported.

Further sampling of basal Cornalean taxa and sequences of complete coding regions would help establish the timing of these duplication events. Due to this uncertainty, both genealogies (Fig 2 -A and -B) were employed for sequence evolution analyses (see Materials and Methods). Results of phylogenetic analyses also indicated that *CorPI*-A is present in all species of *Cornus* examined, while *CorPI*-B is present only in species from the two major subgroups of *Cornus*, CC and DW, the two subgroups with the least species diversity. Within the *CorPI*-A clade, sequences from each morphological subgroup form monophyletic groups with moderate or low support. One of the sequence types (02-08-17) from *C. disciflora*, the only species from the BB clade that does not produce petaloid bracts, appears as a particularly long branch on the tree (Fig 2).

Phylogenetic analyses including both coding and non-coding regions revealed sequence relationships within *CorPI*-A that were strongly supported and concordant in both the Bayesian and NJ analyses (Fig 3). Subgroup relationships within the *CorPI*-A tree are congruent with a previously published phylogeny of the genus *Cornus* inferred from nuclear and chloroplast genes and regions (Fan and Xiang 2001, 2003; Xiang et al. 2002, 2006). In these trees, the DW clade and the BB clade, both having petaloid bracts, form a monophyletic group sister to the CC clade, which has modified, but non-petaloid, bracts. Species relationships within each subgroup are also congruent with the recent hypotheses of *Cornus* phylogeny (Xiang et al. 2006). This genealogical pattern indicates that *CorPI*-A also tracks the phylogeny of *Cornus*.

Multiple sequence types of *CorPI*-A with moderate variations were detected for most species of *Cornus*. In the BW group, sequences form two subclades each of which contains *C. alba*, *C. walteri*, and *C. alsophila*. All of these species have been placed in the subgenus *Kraniopsis* (Xiang and Boufford 2005), suggesting a gene duplication of *CorPI*-A in the common ancestor of this group. The genealogy suggests that this gene duplication in subg. *Kraniopsis* occurred before the divergence between subgenus *Kraniopsis* and subgenus *Mesomora* (Xiang and Boufford 2005), followed by a loss of one copy in subgenus *Mesomora* (Fig 3). Differing sequence types of *CorPI*-A from the same species also group together, suggesting that independent multiple gene duplications occurred within species. At least 3 different sequence types were observed in *C. florida*, 3 in *C. disciflora*, 3 in *C. angustata*, 2 in *C. canadensis*, 2 in *C. controversa* and 3 in *C. oblonga* (Fig 3 and 4). Sequence types detected in some of these species show not only nucleotide substitution differences but also length variance in intron regions, eg. in *C. canadensis*, *C. florida*, *C. disciflora*, and *C. angustata* (Fig 4), suggesting that they are true gene copies instead of allele differences. For example, the long (e.g., Ccana04-01-43) and short (e.g., Ccana04-01-44) copies found in *C. canadensis* differ by an 80 bp indel in intron 1. Similarly, in *C. florida*, the long copy (e.g., clone 01-1-27) is ~1260 bp and short copy (e.g., clone 01-1-2) is ~1130 bp. The two copies are quite divergent, as is shown by the large branch length differences in the gene genealogy (Fig 3). Clone 01-1-15 and clone 01-148-7 found in *C. florida* showing an intermediate relationship to the long and short copies of this species (Fig. 3) are possible resulting from intergenic recombination between the *CorPI* of long and short copies of *C. florida* instead of PCR error (see discussion). In *C. disciflora*, the short copies (e.g., clones 02-08-17, 02-08-23) are ~1200 bp and the long copy (e.g., clone 02-08-2) is ~1260 bp, with

most of the sequence variation accumulated in the 5' end of the sampled region. For clone 02-08-17, the region between the end of the first intron and the beginning portion of exon 2 is highly divergent from other sequences with low homology, suggesting an accelerated rate of sequence evolution or a genomic interruption, which might be the cause of the unusually high level of sequence divergence in this region. The high level of sequence variation observed in the I-domain of this copy, part of which corresponds to the exon 2, probably implicates a loss of function or the origin of a new function for this gene copy. Three sequence variants found in *C. angustata* differ in the level ofsequence variation, but also differ in the observed number of tandem repeats in the first intron region.

Phylogenetic analyses of *CorPI*-B that included both intron and exon regions (results not shown) revealed well supported relationships identical to those found in analyses of exon sequences alone (Fig 2). All three species sampled for CC and two out of three species of DW (*C. canadensis* lacks *CorPI*-B) retain the *CorPI*-B copy and form monophyletic groups, respectively. Multiple sequence types of *CorPI*-B only were detected in *C. mas* (Fig 2). Based on the species phylogeny (((BB, DW), CC), BW), the absence of the *CorPI*-B copy in the BB and BW groups suggests independent loss of this copy in these two lineages. We screened a large number of clones for each sample, analyzed multiple individuals of some species and several species for each subgroup to ensure recovery of all sequence types. However, it is still possible that *CorPI*-B was undetected in BB and BW lineages due to an unpredicted PCR bias.

Rate of substitution, and strength of selection

Pairwise synonymous substitution rates (dS) range from 0 to 0.47 with a mean value of 0.13 in all *CorPI* and are higher than pairwise nonsynonymous substitution rates (dN), which range from 0 to 0.14, with an average of 0.05. The mean dN/dS value among orthologues of *CorPI*-A is greater than that of *CorPI*-B in *Cornus* (0.51 ± 0.42 vs 0.22 ± 0.13, respectively) (Fig. 6). The mean dN/dS for all *CorPI* genes is 0.49 ± 0.38, which is close to the dN/dS value of *CorPI*-A. Among the four morphological subgroups, the mean dN/dS values of *CorPI*-A vary from 1.01 ± 0.50 (in the DW) to 0.57 ± 0.40 (in the CC), with those for the BB and BW groups being intermediate (0.85 ± 0.77 and 0.67 ± 0.48, respectively) (Fig. 6). For the outgroups, the dS value ranges from 0.02 to 0.43, with a mean of 0.27, and the dN ranges from 0.01 to 0.05, with a mean of 0.03. The mean dN/dS value in *Davidia* and *Alangium* is 0.13 ± 0.10, which is smaller than in *Cornus*. The greater dN/dS value for the *CorPI* than *PI*-like genes in outgroups and *CorPI*-A than *CorPI*-B are statistically significant and can not be explained by chance, as assessed by a 2-sample bootstrap resampling test ($P<0.001$). Higher dN/dS values for the *CorPI*-A loci were also found in the BB and DW groups and these were also significantly greater than those found in the BW group($P<0.001$). The CC were found to have the lowest values of dN/dS among the four subgroups, however, a statistical test indicated nosignificant difference of its dN/dS value from those of the BB and DW ($P = 0.27$). This may be the result of fewer data points from the CC group being available for the test.

Relative rate tests (Tajima1993) indicate no significant increases in substitution rate at third codon positions in any of the four subgroups or in either of the two ancient paralog

genes in *Cornus* (Tajima's test, P>0.1). This finding suggests that synonymous substitution rate is homogenous across lineages. Pairwise comparison of dN/dS values for all *CorPI* genes yields values below 1 (94.6%, dN/dS = 0 ~ 1), except for a few pairs of sequences with values slightly above 1 (5.4%, dN/dS = 1 ~ 4) (Fig. 6, *Cornus CorPI*). Plots of dN/dS versus dS indicate that the ratio of dN/dS is negatively related to dS for all categories analyzed (Fig. 7), suggesting that the *CorPI* gene is under strong functional constraints and the rate of nonsynonymous change is constrained as sequences diverge. For *CorPI*-B, although dN/dS and dS show a negative relationship, but the slope is near zero, indicating that dN and dS increase at similar rates as sequences diverge.Thus, the dN/dS value remains unchanged as dS increases. This observation for *CorPI*-B, however, might be an artifact of limited data points.

Codon-based likelihood analyses (Yang 1997) indicate that in the 204 bp coding region, 40 sites (19.6%) are synonymous substitution sites and 164 sites (80.4%) are nonsynonymous sites. If estimates are allowed to change freely along all branches, the dS ratios among the branches vary from 0 to 0.26, the dN values range from 0 to 0.10, and the dN/dS ratios vary from 0 to 0.69 excluding the branches with dS value of zero (approximately one third of the branches have an extremely large omega value due to the zero value of dS; see supplementary materials). Thirty four percent (45 out of 133) of the branches have a dN/dS value over 1, and the maximum dN/dS value is 999 and the minimum is 0.000l (see supplementary materials). Five branches were found having ≥ 2 substitutions at nonsynonymous sites and zero at synonymous sites. These included one of the paralogous genes of *CorPI*-A in *C. florida* (2.0/0), *C. angustata* 02-46-71 (4.1/0), *C. suecica* 27-2-69

(2.5/0), on the branch of *C. oblonga* (2.0/0), and on the branch leading to *CorPI*-B (4.0/0) (Fig 8, and supplementary materials). These data suggest that functional constraints on the gene may restrict substitutions at nonsynonymous sites, but strong positive selection may act at some sites in certain lineages, resulting in large observed dN/dS values.

Under a model allowing differing dN/dS ratios in *CorPI*-A, *CorPI*-B and outgroup species, the estimation of selection on *CorPI*-A and *CorPI*-B are four and five times greater than in the outgroup species (0.50 and 0.35 vs 0.09) (Table 2, Fig 8). Likelihood ratio tests under four of the five models show statistical significance of the difference (Table 3). Under model IV, which assumes an equal ratio between the outgroup species and *Cornus CorPI*-A, an unrealistic model for this test, there is no significant difference between the strength of selection in the *CorPI* and in the outgroup species (Table 3, Yang 1998). The selection force on *CorPI*-A (0.5) is 1.4 times greater than that for *CorPI*-B (0.35), but the likelihood ratio tests under different models (data not shown) did not show significantly statistical support between them (data not shown). Comparison of dN/dS values among the four morphological subgroups of *Cornus* indicate substantial differences in the selection among the groups, with the highest dN/dS value in the DW group (0.70), the only herbaceous group of *Cornus*, the smallest ratio in the CC group (0.34), and intermediate ratios in the BB and BW groups (0.50 and 0.52, respectively) (Fig 8). These estimates are concordant with those estimated from the pair-wise comparison method (Fig. 6).

Under the various codon-based substitution models for testing positive selection at amino acid sites in PAML(Yang et al., 2000), estimates of average dN/dS values range from

0.27 to 0.52 among all models (Table 4). The average dN/dS ratio is less than one indicating that purifying selection dominates the evolution of the gene. This result is congruent with results fromMEGA analysis. The one-ratio model (M0) is easily rejected when compared with all other models allowing the dN/dS ratio to vary among sites. Models that allow for the presence of positively selected sites, that is, M2 (selection), M3 (discrete), and M8 (beta & ω), all suggest the presence of positively selected sites (Table 4). The selection model (M2) suggests ~1.5% of the sites are under positive selection with $\omega_2 = 10.72$. Model 3 (discrete) suggests a large proportion of sites (8.5%) under weak diversifying selection with $\omega_1 = 1.62$ and a small proportion of sites (1.5%) under strong diversifying selection with $\omega_2 = 12.12$. Both models have significantly higher likelihood values than models M0 and M1, providing strong evidence for adaptive evolution (Table 5). Similarly, M8 (beta & ω) suggests ~ 6.7% of sites under positive selection with $\omega = 2.69$. The LRT statistic for comparing M7 (beta) and M8 (beta & ω) indicates that the M8 is significantly better than M7 (Table 5). Amino acid site I_4, the fourth site of I-domain, is identified to be under positive selection at the 99% level by all models that allow for positive selection. At the 95% level, models M2, M3 and M8 also suggested three other positively selected sites: M_{54}, I_6, and I_{11} (Table 4).

Absolute rates and divergence time estimation

Absolute substitution rates of the *CorPI*-A copy estimated using the PL method in r8s 1.71 (Sanderson 2002) for the entire gene, including both of the coding and intron regions, reveal substantial differences among the four major lineages of the genus. When the value of the smoothing parameter is 1, we obtain the best cross-validation score for our data, suggesting that the model deviates from clock-like evolution withthat substantial rate

variation (Sanderson 2002). The substitution rate increases in the subgroups DW and BB (with showy bracts), and DW was found to have the highest rate of substitution (substitutions/site/million year) $3.9*10^{-3} \sim 4.2*10^{-3}$, followed by BB, BW and CC, respectively (Figure 9, and supplementary material). A similar trend is shown by the dN/dS ratio data from PAML analyses. It must be noted that these values may be overestimated given that the calibration dates are based on minimum hypothesized ages for fossils in its representative lineage (Xiang et al. 2005).

Divergence time estimates indicate that the origin and early divergence of the major lineages of *Cornus* occurred during the late Cretaceous and early Tertiary, around 48-90 million years ago (mya). The divergence of *C. oblonga* and of *C. controversa* (52 and 40 mya, respectively), and between *C. chinensis* and *C. officinalis* (39 mya) took place earlier than other speciation events in *Cornus* (Figure 9). For the gene duplication events in *CorPI*-A lineage, the major duplication in the BW group was estimated to be around 44 mya in the Eocene, associated with the divergence of the BW group, while independent gene duplication events within species all occurred after the Oligocene (< 24 mya). For example, gene duplication within the BB group is estimated to have occurred around 12~21 mya in the Miocene and the earliest duplication event in the DW group is placed as a very recent event (< 3 mya, Fig. 9 and supplementary material).

Discussion

Gene polymorphism, duplication, and loss of *CorPI* in *Cornus* –dynamics in time, possible mechanism and evolutionary consequences

A major discovery of this study is the existence of multiple genomic DNA copies and a dynamic gene duplication/loss history of the *PI*-like gene in species of *Cornus*. The sequence variants we detected in each species differ not only at nucleotide sites at a percentage greater than expected from the Taq error (0.1%, Cline et al. 1996), but also in sequence length of intron regions (Fig.4). Thus it is reasonable to believe that the sequence variants represent the true diversity of the gene. Although we do not have data from southern hybridization studies to confirm copy numbers examined in each species in this study, the observation from sequencing multiple individuals for some species supports the existence of multiple copies of the gene in *Cornus* (Fig.4). This level of diversity of *PI*-like genes observed in *Cornus* is significantly higher than what has been known from the mRNA data. Unpublished data (from Dr. Jer-ming Hu) showed only a single copy of *PI* from *C. florida*, a BB species, and two copies from *C. alba*, a BW species. We detected several additional copies of the *PI* gene from genomic DNA sequences in these species, suggesting a more complex evolutionary history of the *PI*-like gene than that inferred from studies of mRNA alone. This finding indicates the importance of examining both mRNA and genomic DNA sequences as a step toward improving our understanding of the evolutionary histories of genes.

The gene genealogy of *PI* homologs suggests that the first and a major duplication event of *CorPI* to form *CorPI*-A and *CorPI*-B likely occurred in the ancestor of *Cornus*, before its radiation into subgroups (Fig. 2A). However, the possibility that this duplication

took place at a deeper node of the Cornales phylogeny, before the divergence of *Cornus* from *Alangium* and *Davidia*, cannot be rejected, especially given that the *PI* homolog from *Davidia* was sometimes grouped with *CorPI*-A, while the *PI*-homolog from *Alangium* was basal to *CorPI*-A, *Davidia,* and *CorPI*-B (Fig. 2-B). After the establishment of *CorPI*-A and *CorPI*-B, the two paralog genes experienced different evolutionary fates. *CorPI*-A was retained in all major lineages of *Cornus* and experienced repeated gene duplication in different lineages at different times, from the middle Tertiary in the BW ancestors (~ 44 mya) to the late Miocene in individual species (e.g., in *C. canadensis* < 3 mya, *C. controversa* < 9 mya, *C. oblonga* < 5 mya, Fig. 9). In contrast, *CorPI*-B copy experienced independent losses in two major subgroups BB and BW and fewer gene duplication events in the groups retaining the B copy (CC and DW). Similar to findings from studies of *PI*-like gene evolution in basal eudicots and basal angiosperms (Kramer et al. 2003; Stellari et al. 2004), we detected multiple copies of the gene in each species studied with most of them derived after the divergence of the species. This suggests frequent gene duplication independently occurring in different species. It is unclear why these species retain multiple copies of the *PI*-like genes. One explanation is that the duplication of the gene may promote genetic divergence during speciation (Irish and Litt 2005). Gene duplication has been regarded as one of the major evolutionary processes generating raw materials for genetic variation, functional diversification, and speciation (Stebbins 1966; Ohno 1970; Levin 1983; Lynch and Conery 2000).

Gene duplications are especially prevalent in plants. The origin of a majority of extant angiosperm lineages are considered to have involved past segmental or whole-genome

duplication events followed by subsequent loss, mutation or epimutation of duplicate gene copies (Wendel 2000). Several mechanisms can result in gene duplication, including genome duplication and chromosomal or segmental duplication. Since all species of *Cornus* are diploids (2n = 18, 20, 22) except for *C. unalaschkensis* (4n = 44) from the DW lineage, and numbers of copies of the gene do not correlate with the chromosome numbers in a species, the multiple copies of the *CorPI* gene were likely derived from segmental duplications. Genome duplication and chromosomal duplication may not play a role in the presence of multiple copies of the gene in the genus, even in the 4n species *C. unalaschkensis* from the DW group. The sample analyzed for the 4n species does not combine genes from its two putative 2n parental species, *C. suecica* and *C. canadensis* (Figs. 2. 3), suggesting gene loss from one of the parental species. The *CorPI* gene diversity found in species of *Cornus* is mostly the result of gene duplication. Other mechanisms, however, such as gene recombination may also have contributed to a smaller extent to the gene diversity observed. For example, in two different individuals of *C. florida*, two recombinant sequences with the same split point were found, implying a sequence evolving from intergenic recombination in this species.

Duplicated gene copies are traditionally known to have two evolutionary fates; the origin of a new gene with new function (neofunctionalization), or becoming a pseudogene by losing function. Recent studies have demonstrated that duplicated gene copies can also be redundant in function by diverging in their roles to retain different subfunctions of the original gene (Force et al. 1999; Lynch and Conery 2000). The maintenance of redundant duplicated genes can contribute to the robustness of a genetic network by reducing the fitness

effects of deleterious mutations (Gu 2003). Recent studies of B-class genes in lower eudicots and basal angiosperms revealed that the paralogous genes display divergent expression patterns, supporting sub/neofunction of the duplicated genes and their potential roles in the evolution of new floral phenotypes (Purugganan et al. 1998; Kramer et al. 1999; Di Stilio et al. 2005). In *Cornus*, functional studies are necessary to test the role and genetic consequence of maintaining multiple copies of *CorPI* paralogs in the species. By now, we observed no direct correlation between the duplication pattern of *CorPI* evolution and petaloid bracts in the BB and DW subgroups.

Selection on *CorPI* in *Cornus*

The mean dN/dS value of *CorPI* is 0.49 ± 0.38 (Fig. 6) and 0.47 (Table 2, model C) estimated based on the pairwise comparison and codon-based likelihood methods, respectively. These values are five times larger than the dN/dS value detected in the outgroups (*Davidia* and *Alangium*) 0.13 ± 0.10 (see results, pairwise comparison) and 0.09 (Table 2).These values are comparable to the value of 0.11-0.19 reported for the MADS-box floral homeotic genes found in other species (Purugganan et al. 1995). Besides the elevated dN/dS ratio of *CorPI* observed in this study, analysis of the anthocyanin regulatory gene in *Cornus* also revealed a dN/dS ratio greater than 0.4 (Fan et al. 2004). These dN/dS values for the regulatory genes observed in *Cornus* are generally larger than the mean dN/dS value of 0.14 detected for other plant nuclear loci, suggesting that relaxation in selection constraint of these regulatory genes may have provided the genetic basis for the evolution of the striking heterogeneous inflorescence and bract morphology, and the various fruit color found in this genus. After gene duplication of *CorPI* into *CorPI*-A and *CorPI*-B, the functional constraints

were relaxed in both paralogous genes as compared to the outgroup genera. Compared to the *CorPI*-A and *CorPI*-B copies found in *Cornus*, the dN/dS ratio in *CorPI*-A is about two times greater than *CorPI*-B (0.51 ± 0.42 vs 0.22 ± 0.13 in pairwise comparison, Fig 6; 0.50 vs 0.35 in PAML, Fig 8), although the difference of the ratios between the A and B copies was tested to be non-significant in PAML analyses. The higher dN/dS ratio of *CorPI*-A can be the result of relaxation of its original functional constraint, which may be associated with gene duplication and long-term maintenance of older paralogous gene copies. Our data indicate that one or both of the more ancient paralog copies may have experienced diversifying selection, probably for a subfunction or neofunction, based on analysis of amino acid sites (see below and table 4).

An elevated dN/dS ratio was observed for the subgroups with petaloid bracts, DW and the BB (the mean dN/dS ratios of different morphological subgroups show a trend of DW > (BB, BW) > CC (Fig. 6 and 8). Relative rate tests (Tajima 1993) indicate that these lineages (DW and BB) do not display a significant increase in nucleotide substitution rates for the largely synonymous third codon positions, compared to the CC and BW lineages (Tajima's test, $P > 0.1$), suggesting the increase of dN/dS ratios in DW and BB was not due to an increase in synonymous substitution rates. This further showed that the *CorPI* experienced diversifying selection in the DW and BB groups, which might be due to the accelerated nonsynonymous substitution rate of the lineages. However, it cannot be determined whether the elevated dN/dS values in BB and DW are associated with the evolution of the petaloid bracts in these two subgroups.

Our data reveal an uneven distribution of selection pressure among amino acid sites on the *CorPI* gene (Table 4 and 5). Four amino acid sites were detected under diversifying positive selection, and three of these sites are located in the I-domain, while most amino acid sites were under purifying selection (Table 4, Fig. 5). The substitution of these positively selected amino acid sites involved changes of polarity and charge of the amino acids. At sites I_4 and I_{11}, the amino acid change between polar and nonpolar, and at sites I_6 change among polar, nonpolar and positively charged amino acids (Table 6). At the site M_{54}, the only positive selected site identified on the partially sampled MADS-domain, amino acid substitution only involves changes within nonpolar amino acids. The I-domains of plant MADS-box genes are known to evolve rapidly (Martinez-Castilla and Alvarez-Buylla 2003). They are shown to be associated with dimerizational and functional specificity (Riechmann et al. 1996a; Krizek and Meyerowitz 1996a). The I-domain has the critical function of interaction between functional counterparts *AP3* and *PI* to form heterodimers in order to execute its function in petal and stamen development in eudicots (Winter et al. 2002). Therefore, amino acid changes detected in the I-domain of *CorPI* genes may have brought profound change or modification to its regulatory function. The simulation study showed that the power of detecting adaptive evolution in genes (positive selection) increases in longer sequences and the power for detecting short sequences is almost 0%, unless selection is strong (Anisimova et al. 2001). The finding of positive selection in only 204 bp of *PI* we examined indicates strong diversifying selection at these sites and likely reflects a role in functional diversity of the gene. Mapping of these sites onto the phylogeny indicates that the changes at the three sites detected on the I-domain centered in the major lineages of *Cornus* on the basal nodes, suggesting that positive selection at these sites occurred when the genus

started its early radiation (Figs of supplementary materials). Recent functional studies also show that there is no relationship between expression divergence and protein divergence (Wagner 2000; Gu et al. 2002; Makova and Li 2003; Duarte et al. 2006) suggesting that it is risky to infer functional divergence based only on protein sequence changes. Thus, although positive selection detected for the gene (Table 4 and discussion below) suggests functional divergence of the paralogous copies, functional studies are necessary to test the hypothesis.

Even though diversifying selection at amino acid sites of the I-domain is present, the average dN/dS ratio of the *CorPI* sequence examined is well below one and the dS values and dN/dS ratios are negatively related, which commonly indicates purifying selection of the gene under analysis. This raises the need for caution, since using the aggregate dN/dS value of a gene to determine strength of selection can miss important information. The dN/dS ratio measured overof a full gene can be dominated by strong selection on just a few sites. Tests for positive selection at individual sites are therefore valuable for gaining a more complete and accurate picture of the mechanisms underlying sequence evolution. It is known that most proteins are found under purifying selection with a dN/dS ratio close to 0, because the majority of amino acids are i under strong functional selection to remain invariant (Li et al. 1985).

Rate heterogeneity of sequence evolution

Among the major lineages of *Cornus*, rates of substitution in *CorPI*-A are three to ten times higher in DW than in BB, BW and CC (Fig 8). Several biological causes have been

proposed to explain rate heterogeneity of molecular evolution between lineages (also known as lineage effects), including generation time effects (Wu and Li 1985; Bousquet et al. 1992; Gaut et al. 1992; Whittle and Johnston 2002, 2003; Wright, Lauga and Charlesworth 2002), differences in metabolic rates (Martin and Palumbi 1993), unequal efficiency of DNA repair (Britten 1986), and population size (Li 1997). Recently, species diversity was also shown to be positively correlated with substitution rates in flowering plants and other organisms (Barraclough et al. 1998; Barraclough and Savolainen 2001; Webster 2003; Xiang et al. 2004). Because only DW is herbaceous in *Cornus*, while species in other subgroups are small trees, generation time might be an important factor causing the accelerated rate of evolutoin of the *PI* gene in DW. An elevated rate of molecular evolution in DW was also observed in other molecular markers (Fan and Xiang 2001; Xiang et al. 2006), indicating that this pattern could be a genome-wide phenomenon. There is little reason to believe that metabolic rate and DNA repair efficiency differs among the four groups (these two causes are often applied to animals only). Furthermore, there is no evidence indicating major differences in population size among the lineages.

A fast evolving copy of *CorPI*-A was detected in *C. disciflora* (a species without petaloid bracts from the BB group). The accelerated rate of sequence evolution of the species is shown as a long branch connecting the species on the gene tree and is largely due to increased sequence divergence in the I-domain (Fig. 2 and 5, Cdisc 02-08-17). It is possible that this sequence may represent a pseudogene. However, no stop codons were found in the entire region sequenced, and the dN/dS ratio for the branch is well below one (dN/dS = 0.38) and not greater than in other branches of the BB clade (ave dN/dS = 0.50, Fig 8). These data

do not agree with the expectation for a pseudogene (dN/dS close to 1). Detailed examination of this sequence showed that it also has a unique sequence of 80 bp located in the intron region preceding the I-domain (exon 2), as well as several amino acid mutations in the 5' end of the I-domain (Fig 5). It is also possible that a genomic interruption (i.e., exchange of the 80 bp) has caused a dramatic difference in this sequence and also brought changes to the adjacent coding region of this gene copy. Under this scenario mutations in the I-domain of this sequence may have led to functional divergence of this copy in *C. disciflora*, the single species of the big-bracted group (BB) lacking the large bracts. Recent studies have shown that genomic interruption of certain genes could induce novel functions of the gene (Chen et al. 1997). At present, we have no evidence to confirm this hypothesis. Sequencing the full length of this gene copy and a functional study are necessary to determine whether it is in fact a pseudogene or a new gene copy with a possible new function.

Phylogenetic utility

We detected sufficient sequence variation and parsimony informative sites from *CorPI* for a phylogenetic analysis of *Cornus* (see Results above). The species relationships suggested by the *CorPI*-A copy are well supported and highly consistent with those suggested by sequence data from multiple genes from the chloroplast and nuclear genomes (Fan and Xiang 2001, Xiang et al. 2006; Fig. 3). This evidence suggests that the gene copy also tracks the phylogeny of *Cornus* and is useful for phylogenetic analysis. Caution, however, is needed when comparing sequences within the BW lineage due to the existence of two relatively ancient paralogous copies in most species.

Conclusion

In this study, we reveal dynamic history of *PI*-like gene duplication and loss at different periods in the history of *Cornus*. Two ancient paralogs, *CorPI*-A and *CorPI*-B, were detected in *Cornus,* followed by subsequent losses of the B copy in two major subgroups during the early radiation of the genus. Each species analyzed was found to contain multiple copies with most of them derived recently within a species. We found that the selection force (dN/dS ratio) on the gene was more relaxed in the *Cornus* lineage than in the closely related outgroups, *Alangium* and *Davidia*, and in other MADS-box genes reported for other flowering plants. Strong positive selection at four amino acid sites of *CorPI* were also detected, with three of these sites from the I-domain of the gene, a region critical for dimerization activity. Selection on the *CorPI* genes also differs among major copies and among the *Cornus* lineages. The total substitution rates of the *CorPI* gene also differ among lineages, and an elevated rate was observed for the DW lineage, the only herbaceous group of *Cornus*. Finally, we found that the *CorPI*-A copy also tracks the phylogeny of *Cornus*.

Future directions

It is necessary to confirm the copy number of the *CorPI* genes detected in this study by southern blotting analysis and to further test whether these genomic copies are also present in the expressed mRNA. To understand if the paralog genes of *CorPI* are functional redundancy or evolve subfunction or a new function, expression and functional studies need to be carried out. We have also investigated *AP3*-like gene evolution to gain a better picture of B-class gene evolution in *Cornus*. These studies compare the pattern of B-class gene evolution in the

group, providing more insights into B-class gene evolution in eudicot plants and regulatory gene evolution in flowering plants as a whole. Finally, it will be intriguing to test if B-class gene evolution contributed to the evolution of the morphologically novel, petaloid bracts in *Cornus*.

Acknowledgement

We thank Drs. Jer-Ming Hu and Michael Frohlic who provided primers and mRNA sequence data of *C. florida* and *C.alba* for this study. We are also grateful to the NSF supported DEEPTIME program (funded to D. E. Soltis DEB-0090283) for the travel support to workshops on divergence time dating and application of fossil data. We thank M. D. Purugganan, B. M. Wiegmann, and N. S. Allen for the insightful comments on an earlier version of the manuscript. We thank D. Thomas and K. M. Ding for help with collecting data. This study is supported by National Science Foundation grants to Q-Y (J.) X (NSF-DEB 0129069 and DEB 0444125).

References

ALTSCHUL, S. F., T. L. MADDEN, A. A. SCHAFFER, J. ZHANG, Z. ZHANG, W. MILLER, AND D. J. LIPMAN. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25: 3389-3402.

ALVAREZ-BUYLLA, E. R., S. PELAZ, S. J. LILJEGREN, S. E. GOLD, C. BURGEFF, G. S. DITTA, L. R. DE POUPLANA, L. MARTINEZ-CASTILLA, AND M. F. YANOFSKY. 2000. An ancestral MADS-box gene duplication occurred before the

divergence of plants and animals. Proceedings of the National Academy of Sciences of the United States of America 97: 5328-5333.

AMBROSE, B. A., D. R. LERNER, P. CICERI, C. M. PADILLA, M. F. YANOFSKY, AND R. J. SCHMIDT. 2000. Molecular and genetic analyses of the *silky1* gene reveal conservation in floral organ specification between eudicots and monocots. Molecular Cell 5: 569-579.

ANISIMOVA, M., J. P. BIELAWSKI, AND Z. YANG. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Molecular Biology and Evolution 18: 1585-1592.

AOKI, S., K. UEHARA, M. IMAFUKU, M. HASEBE, AND M. ITO. 2004. Phylogeny and divergence of basal angiosperms inferred from *APETALA3*- and *PISTILLATA*-like MADS-box genes. Journal of Plant Research 117: 229-244.

BARRACLOUGH, T. G. AND V. SAVOLAINEN. 2001. Evolutionary rates and species diversity in flowering plants. Evolution 55: 677-683.

BARRACLOUGH, T. G., P. H. VOGLER, AND P. H. HARVEY. 1998. Revealing the factors that promote speciation. Philosophical Transactions of the Royal Society of London B 353: 241-249.

BARRIER, M., R. H. ROBICHAUX, AND M. D. PURUGGANAN. 2001. Accelerated regulatory gene evolution in an adaptive radiation. Proceedings of the National Academy of Sciences of the United States of America 98: 10208-10213.

BECKER, A., K. U. WINTER, B. MEYER, H. SAEDLER, AND G. THEISSEN. 2000. MADS-box gene diversity in seed plants 300 million years ago. Molecular Biology and Evolution 17: 1425-1434.

BOUSQUET, J., S. H. STRAUSS, A. H. DOERKSEN, AND R. A. PRICE. 1992. Extensive variation in evolutionary rate of *rbcL* gene sequences among seed plants. Proceedings of the National Academy of Sciences of the United States of America 89: 7844-7848.

BRITTEN, R. J. 1986. Rates of DNA sequence evolution differ between taxonomic groups. Science 231: 1393-1398.

CHANG, S. M., Y. LU, AND M. D. RAUSHER. 2005. Neutral evolution of the nonbinding region of the anthocyanin regulatory gene *Ipmyb1* in *Ipomoea*. Genetics 170: 1967-1978.

CHEN, J. J., B. J. JANSSEN, A. WILLIAMS, AND N. SINHA. 1997. A gene fusion at a homeobox locus: alterations in leaf shape and implications for morphological evolution. Plant Cell 9: 1289-1304.

CLINE, J., J. C. BRAMAN, AND H. H. HOGREFE. 1996. PCR fidelity of *pfu* DNA polymerase and other thermostable DNA polymerases. Nucleic Acids Research 24: 3546-3551.

COEN, E. S. AND E. M. MEYEROWITZ. 1991. The war of the whorls: genetic interactions controlling flower development. Nature 353: 31-37.

DI STILIO, V. S., E. M. KRAMER, AND D. A. BAUM. 2005. Floral MADS box genes and homeotic gender dimorphism in *Thalictrum dioicum* (Ranunculaceae) - a new model for the study of dioecy. Plant Journal 41: 755-766.

DOEBLEY, J. AND L. LUKENS. 1998. Transcriptional regulators and the evolution of plant form. Plant Cell 10: 1075-1082.

DUARTE, J. M., L. CUI, P. K. WALL, Q. ZHANG, X. ZHANG, J. LEEBENS-MACK, H. MA, N. ALTMAN, AND C. W. DEPAMPHILIS. 2006. Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of *Arabidopsis*. Molecular Biology and Evolution 23: 469-478.

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

FAN, C., M. D. PURUGGANAN, D. T. THOMAS, B. M. WIEGMANN, AND J. Q. XIANG. 2004. Heterogeneous evolution of the *Myc*-like Anthocyanin regulatory gene and its phylogenetic utility in *Cornus* L. (Cornaceae). Molecular Phylogenetics and Evolution 33: 580-594.

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

-----. 2003. Phylogenetic analyses of Cornales based on 26S rDNA and combined 26S rDNA-*matK*-*rbcL* sequence data. American Journal of Botany 90: 1357-1372.

FARES, M. A., D. BEZEMER, A. MOYA, AND I. MARIN. 2003. Selection on coding regions determined *Hox7* genes evolution. Molecular Biology and Evolution 20: 2104-2112.

FORCE, A., M. LYNCH, F. B. PICKETT, A. AMORES, Y. L. YAN, AND J. POSTLETHWAIT. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151: 1531-1545.

GAUT, B. S., S. V. MUSE, W. D. CLARK, AND M. T. CLEGG. 1992. Relative rates of nucleotide substitution at the *rbcL* locus of monocotyledonous plants. Journal of Molecular Evolution 35: 292-303.

GOLDMAN, N. AND Z. YANG. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular.Biology and Evolution 11: 725-736.

GOTO, K. AND E. M. MEYEROWITZ. 1994. Function and regulation of the *Arabidopsis* floral homeotic gene *PISTILLATA*. Genes and Development 8: 1548-1560.

GU, X. 2003. Evolution of duplicate genes versus genetic robustness against null mutations. Trends in Genetics 19: 354-356.

GU, Z., D. NICOLAE, H. H. LU, AND W. H. LI. 2002. Rapid divergence in expression between duplicate genes inferred from microarray data. Trends in Genetics 18: 609-613.

HUBER, T., G. FAULKNER, AND P. HUGENHOLTZ. 2004. Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. Bioinformatics 20: 2317-2319.

HUELSENBECK, J. P. AND F. RONQUIST. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17: 754-755.

IRISH, V. F. AND A. LITT. 2005. Flower development and evolution: gene duplication, diversification and redeployment. Current Opinion in Genetics and Development 15: 454-460.

JACK, T. 2001. Plant development going MADS. Plant Molecular Biology 46: 515-520.

JACK, T., L. L. BROCKMAN, AND E. M. MEYEROWITZ. 1992. The homeotic gene *APETALA3* of *Arabidopsis thaliana* encodes a MADS box and is expressed in petals and stamens. Cell 68: 683-697.

KIM, S., J. KOH, H. MA, Y. HU, P. K. ENDRESS, B. A. HAUSER, M. BUZGO, P. S. SOLTIS, AND D. E. SOLTIS. 2005a. Sequence and expression studies of A-, B-, and E-class MADS-box homologues in *Eupomatia* (Eupomatiaceae): Support for the bracteate origin of the calyptra. International Journal of Plant Sciences 166: 185-198.

KIM, S., J. KOH, M. J. YOO, H. Z. KONG, Y. HU, H. MA, P. S. SOLTIS, AND D. E. SOLTIS. 2005b. Expression of floral MADS-box genes in basal angiosperms: implications for the evolution of floral regulators. Plant Journal 43: 724-744.

KIM, S. T., M. J. YOO, V. A. ALBERT, J. S. FARRIS, P. S. SOLTIS, AND D. E. SOLTIS. 2004. Phylogeny and diversification of B-function MADS-box genes in angiosperms: Evolutionary and functional implications of a 260-million-year-old duplication. American Journal of Botany 91: 2102-2118.

KRAMER, E. M., V. S. DI STILIO, AND P. M. SCHLUTER. 2003. Complex patterns of gene duplication in the *APETALA3* and *PISTILLATA* lineages of the Ranunculaceae. International Journal of Plant Sciences 164: 1-11.

KRAMER, E. M., R. L. DORIT, AND V. F. IRISH. 1998. Molecular evolution of genes controlling petal and stamen development: duplication and divergence within the *APETALA3* and *PISTILLATA* MADS-box gene lineages. Genetics 149: 765-783.

KRAMER, E. M. AND V. F. IRISH. 1999. Evolution of genetic mechanisms controlling petal development. Nature 399: 144-148.

-----. 2000. Evolution of the petal and stamen developmental programs: Evidence from comparative studies of the lower eudicots and basal angiosperms. International Journal of Plant Sciences 161: S29-S40.

KRAMER, E. M., M. A. JARAMILLO, AND V. S. DI STILIO. 2004. Patterns of gene duplication and functional evolution during the diversification of the *AGAMOUS* subfamily of MADS box genes in angiosperms. Genetics 166: 1011-1023.

KRIZEK, B. A. AND E. M. MEYEROWITZ. 1996a. Mapping the protein regions responsible for the functional specificities of the *Arabidopsis* MADS domain organ-identity proteins. Proceedings of the National Academy of Sciences of the United States of America 93: 4063-4070.

-----. 1996b. The *Arabidopsis* homeotic genes *APETALA3* and *PISTILLATA* are sufficient to provide the B class organ identity function. Development 122: 11-22.

KUMAR, S., K. TAMURA, AND M. NEI. 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Briefings in Bioinformatics 5: 150-163.

LAMB, R. S. AND V. F. IRISH. 2003. Functional divergence within the *APETALA3*/*PISTILLATA* floral homeotic gene lineages. Proceedings of the National Academy of Sciences of the United States of America 100: 6558-6563.

LEVIN, D. A. 1983. Polyploidy and novelty in flowering plants. American Naturalist 122: 1-25.

LI, W. H. 1997. Molecular evolution. Sinauer Associates, Sunderland, Mass.

LI, W. H., C. I. WU, AND C. C. LUO. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Molecular Biology and Evolution 2: 150-174.

LITT, A. AND V. F. IRISH. 2003. Duplication and diversification in the *APETALA1/FRUITFULL* floral homeotic gene lineage: Implications for the evolution of floral development. Genetics 165: 821-833.

LOWE, C. J. AND G. A. WRAY. 1997. Radical alterations in the roles of homeobox genes during echinoderm evolution. Nature 389: 718-721.

LYNCH, M. AND J. S. CONERY. 2000. The evolutionary fate and consequences of duplicate genes. Science 290: 1151-1155.

MAKOVA, K. D. AND W. H. LI. 2003. Divergence in the spatial pattern of gene expression between human duplicate genes. Genome Research 13: 1638-1645.

MARTIN, A. P. AND S. R. PALUMBI. 1993. Body size, metabolic rate, generation time, and the molecular clock. Proceedings of the National Academy of Sciences of the United States of America 90: 4087-4091.

MARTINEZ-CASTILLA, L. P. AND E. R. ALVAREZ-BUYLLA. 2003. Adaptive evolution in the *Arabidopsis* MADS-box gene family inferred from its complete resolved phylogeny. Proceedings of the National Academy of Sciences of the United States of America 100: 13407-13412.

NAM, J., J. KIM, S. LEE, G. H. AN, H. MA, AND M. S. NEI. 2004. Type I MADS-box genes have experienced faster birth-and-death evolution than type II MADS-box genes in angiosperms. Proceedings of the National Academy of Sciences of the United States of America 101: 1910-1915.

NEI, M. AND T. GOJOBORI. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular Biology and Evolution 3: 418-426.

NG, M. AND M. F. YANOFSKY. 2001. Function and evolution of the plant MADS-box gene family. Nature Reviews Genetics 2: 186-195.

OHNO S 1970. Evolution by gene duplication. Springer, New York.

PARENICOVA, L., S. DE FOLTER, M. KIEFFER, D. S. HORNER, C. FAVALLI, J. BUSSCHER, H. E. COOK, R. M. INGRAM, M. M. KATER, B. DAVIES, G. C. ANGENENT, AND L. COLOMBO. 2003. Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in *Arabidopsis*: new openings to the MADS world. Plant Cell 15: 1538-1551.

PELAZ, S., G. S. DITTA, E. BAUMANN, E. WISMAN, AND M. F. YANOFSKY. 2000. B and C floral organ identity functions require *SEPALLATA* MADS-box genes. Nature 405: 200-203.

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

PURUGGANAN, M. D. 1998. The molecular evolution of development. Bioessays 20: 700-711.

PURUGGANAN, M. D., S. D. ROUNSLEY, R. J. SCHMIDT, AND M. F. YANOFSKY. 1995. Molecular evolution of flower development: diversification of the plant MADS-box regulatory gene family. Genetics 140: 345-356.

RAMBAUT, A. AND A. J. DRUMMOND 2004. Tracer. University of Oxford, Oxford.

REMINGTON, D. L. AND M. D. PURUGGANAN. 2002. *GAI* homologues in the Hawaiian silversword alliance (Asteraceae-Madiinae): molecular evolution of growth regulators in a rapidly diversifying plant lineage. Molecular Biology and Evolution 19: 1563-1574.

RIECHMANN, J. L., B. A. KRIZEK, AND E. M. MEYEROWITZ. 1996a. Dimerization specificity of Arabidopsis MADS domain homeotic proteins *APETALA1*, *APETALA3*, *PISTILLATA*, and *AGAMOUS*. Proceedings of the National Academy of Sciences of the United States of America 93: 4793-4798.

RIECHMANN, J. L. AND E. M. MEYEROWITZ. 1997. MADS domain proteins in plant development. Biological Chemistry 378: 1079-1101.

RIECHMANN, J. L., M. WANG, AND E. M. MEYEROWITZ. 1996b. DNA-binding properties of *Arabidopsis* MADS domain homeotic proteins *APETALA1*, *APETALA3*, *PISTILLATA* and *AGAMOUS*. Nucleic Acids Research 24: 3134-3141.

RONQUIST, F. AND J. P. HUELSENBECK. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572-1574.

SAITOU, N. AND M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4: 406-425.

SANDERSON, M. J. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Molecular Biology and Evolution 19: 101-109.

SCHWARZ-SOMMER, Z., I. HUE, P. HUIJSER, P. J. FLOR, R. HANSEN, F. TETENS, W. E. LONNIG, H. SAEDLER, AND H. SOMMER. 1992. Characterization of the *Antirrhinum* floral homeotic MADS-box gene deficiens: evidence for DNA binding and autoregulation of its persistent expression throughout flower development. EMBO Journal 11: 251-263.

SCHWARZ-SOMMER, Z., P. HUIJSER, W. NACKEN, H. SAEDLER, AND H. SOMMER. 1990. Genetic control of flower development by homeotic genes in *Antirrhinum majus*. Sciene 250: 931-936.

STEBBINS, G. L. 1966. Processes of Organic Evolution. Englewood Cliffs, N.J., Prentice-Hall.

STELLARI, G. M., M. A. JARAMILLO, AND E. M. KRAMER. 2004. Evolution of the *APETALA3* and *PISTILLATA* lineages of MADS-box-containing genes in the basal angiosperms. Molecular Biology and Evolution 21: 506-519.

SWOFFORD, D. L. 2002. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4.10b. Sinauer Associates, Sunderland, Mass.

TAJIMA, F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. Genetics 135: 599-607.

THEISSEN, G. 2001. Development of floral organ identity: stories from the MADS house. Current Opinion in Plant Biology 4: 75-85.

THEISSEN, G., A. BECKER, A. DI ROSA, A. KANNO, J. T. KIM, T. MUNSTER, K. U. WINTER, AND H. SAEDLER. 2000. A short history of MADS-box genes in plants. Plant Molecular Biology 42: 115-149.

THEISSEN, G. AND H. SAEDLER. 2001. Plant biology. Floral quartets. Nature 409: 469-471.

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 25: 4876-4882.

TROBNER, W., L. RAMIREZ, P. MOTTE, I. HUE, P. HUIJSER, W. E. LONNIG, H. SAEDLER, H. SOMMER, AND Z. SCHWARZ-SOMMER. 1992. *GLOBOSA*: a homeotic gene which interacts with *DEFICIENS* in the control of *Antirrhinum* floral organogenesis. EMBO Journal 11: 4693-4704.

VAN DE, P. Y., J. S. TAYLOR, I. BRAASCH, AND A. MEYER. 2001. The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. Journal of Molecular Evolution 53: 436-446.

VANDENBUSSCHE, M., J. ZETHOF, S. ROYAERT, K. WETERINGS, AND T. GERATS. 2004. The duplicated B-class heterodimer model: whorl-specific effects and complex genetic interactions in *Petunia hybrida* flower development. Plant Cell 16: 741-754.

WAGNER, A. 2000. Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate. Proceedings of the National Academy of Sciences of the United States of America 97: 6579-6584.

WEBSTER, A. J., R. J. PAYNE, AND M. PAGEL. 2003. Molecular phylogenies link rates of evolution and speciation. Science 301: 478.

WEIGEL, D. AND E. M. MEYEROWITZ. 1994. The ABCs of floral homeotic genes. Cell 78: 203-209.

WENDEL, J. F. 2000. Genome evolution in polyploids. Plant Molecular Biology 42: 225-249.

WHIPPLE, C. J., P. CICERI, C. M. PADILLA, B. A. AMBROSE, S. L. BANDONG, AND R. J. SCHMIDT. 2004. Conservation of B-class floral homeotic gene function between maize and *Arabidopsis*. Development 131: 6083-6091.

WHITTLE, C. A. AND M. O. JOHNSTON. 2002. Male-driven evolution of mitochondrial and chloroplastidial DNA sequences in plants. Molecular Biology and Evolution 19: 938-949.

-----. 2003. Broad-scale analysis contradicts the theory that generation time affects molecular evolutionary rates in plants. Journal of Molecular Evolution 56: 223-233.

WINTER, K. U., C.WEISER, K. KAUFMANN, A. BOHNE, C. KIRCHNER, A. KANNO, H. SAEDLER, AND G. THEISSEN. 2002. Evolution of class B floral homeotic proteins: obligate heterodimerization originated from homodimerization. Molecular Biology and Evolution 19: 587-596.

WRIGHT, S. I., B. LAUGA, AND D. CHARLESWORTH. 2002. Rates and patterns of molecular evolution in inbred and outbred *Arabidopsis*. Molecular Biology and Evolution 19: 1407-1420.

WU, C. I. AND W. H. LI. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. Proceedings of the National Academy of Sciences of the United States of America 82: 1741-1745.

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

XIANG, Q. Y., D. T. THOMAS, W. ZHANG, S. R. MANCHESTER, AND Z. MURRELL. 2006. Species level phylogeny of the genus *Cornus* (Cornaceae) based on molecular and morphological evidence - implications for taxonomy and Tertiary intercontinental migration. Taxon 55: 9-30.

XIANG, J. Q., M. L. 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, J. Q., 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, J. Q., 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.

XIANG, Q. Y. AND D. E. BOUFFORD. 2005. Cornaceae, Mastixiaceae, Toricelliaceae, Helwingiacaee, Aucubaceae. In Z. Y. Wu and P. H. Raven [eds.], *Flora of China* (Apiaceae through Ericaceae), vol. 14, 206-234. Science Press, Beijing, and Missouri Botanical Garden Press, St. Louis.

XIANG, Q. Y., S. R. MANCHESTER, D. T. THOMAS, W. ZHANG, AND C. FAN. 2005. Phylogeny, biogeography, and molecular dating of cornelian cherries (*Cornus*, Cornaceae): tracking Tertiary plant migration. Evolution 59: 1685-1700.

XIANG, Q. Y., W. H. ZHANG, R. E. RICKLEFS, H. QIAN, Z. D. CHEN, J. WEN, AND J. L. HUA. 2004. Regional differences in rates of plant speciation and molecular evolution: a comparison between eastern Asia and eastern North America. Evolution 58: 2175-2184.

YANG, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Computer applications in the biosciences 13: 555-556.

-----. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Molecular Biology and Evolution 15: 568-573.

YANG, Z. AND R. NIELSEN. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Molecular Biology and Evolution 19: 908-917.

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.

YANG, Z., W. S. WONG, AND R. NIELSEN. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. Molecular Biology and Evolution 22: 1107-1118.

ZAHN, L. M., H. KONG, J. H. LEEBENS-MACK, S. KIM, P. S. SOLTIS, L. L. LANDHERR, D. E. SOLTIS, C. W. DEPAMPHILIS, AND H. MA. 2005. The evolution of the *SEPALLATA* subfamily of MADS-box genes: a preangiosperm origin with multiple duplications throughout angiosperm history. Genetics 169: 2209-2223.

Table 1. Source of plant materials used in this study and information of clones analyzed. Specimens used in this study including taxa represented, reference identifiers, vouchers. Voucher specimens were collected by Q.Y. (J.) Xiang and deposited in NCSC.

Subgroups†	Species	Voucher and collection locality	Abbreviation in analyses	No. of clones analyzed	No. of clones sequenced
BW	*C. alba* L.	02-269, Beijing China	Calba02-269	10	5
	C. alsophila W.W.Sm.	02-142, Yunnan, China	Calso02-142	10	7
	C. alsophila W.W.Sm.	02-206, Yunnan, China	Calso02-206	40	7
	C. alternifolia L.	04-06, West Virginia, USA	Calte04-06	40	6
	C. controversa Hemsl.	02-94, Sichuan, China	Ccont02-94	49	9
	C. oblonga Wall.	02-223, Yunnan, China	Coblo02-223	63	11
	C. walteri Wangerin	02-267, Beijing, China	Cwalt02-267	12	8
	C. walteri Wangerin	02-161, Yunnan, China	Cwalt02-161	39	6
CC	*C. chinensis* Wangerin	02-83, Sichuan, China	Cchin02-83	15	6
	C. mas L.	02-282, JC Raulston Arboretum	Cmas02-282	39	11
	C. officinalis Siebold & Zucc.	918-85A, Arnold Arboretum	Coffi918-85A	48	10

Table 1. continued.

Subgroups†	Species	Voucher and collection locality	Abbreviation in analyses	No. of clones analyzed	No. of clones sequenced
BB	*C. angustata* (Chun)T.R.Dudley	02-46, Guangxi, China	Cangu02-46	39	9
	C. disciflora Sesse & Moc. ex DC.	02-08, Heredia, Costa Rico	Cdisc02-08	49	11
	C. florida L.	01-1, North Carolina, USA	Cflor01-1	62	14
	C. florida L.	01-148, North Carolina, USA	Cflor01-148	39	11
	C. florida L.	02-16, Javier Clauijero, Costa Rico	Cflor02-16	14	10
	C. kousa Buerger ex Miq.	02-268, Beijing, China	Ckous02-268	47	11
	C. kousa Buerger ex Miq.	04-C45, Sichuan, China	Ckous04-C45	39	13
DW	*C. canadensis* L.	Pop6-1, British Columbia, Canada	Ccana6-1	41	8
	C. canadensis L.	04-01, West Virginia, USA	Ccana04-01	40	13
	C. suecica L.	Pop27-2, Alaska, USA	Csuec27-2	13	9
	C. unalaschkensis Ledeb.	Pop1-1, Washington, USA	Cunal1-1	40	10
Outgroups	*Alangium sp.*	02-72, Guangxi, China	Alan02-72	22	10
	Davidia involucrate Baill.	02-87, Sichuan, China	Davi02-87	23	10

†BW = blue- or white- fruited group; CC = cornelian cherry; BB = big-bracted group; DW = dwarf dogwoods.

Table 2. Various analyses which allow different dN/dS ratios among *CorPI*-A, *CorPI*-B and outgroups to detect differences of selection forces (ω = dN/dS) between *CorPI* of *Cornus* and *PI*-like genes in outgroups using PAML 3.15. Log likelihood values and parameters were estimated for each model analyzed.

Model	p	l	K	ω_0	ω_1	ω_2
A. Three ratio: ω0, ω1, ω2	3	-1357.72	1.97	0.094	0.496	0.351
B. One ratio: ω0 = ω1 =ω2	1	-1363.88	1.97	0.375	$=\omega_0$	$=\omega_0$
C. Two ratio: ω0, ω1=ω2	2	-1358.44	2.00	0.094	0.468	$=\omega_1$
D. Two ratio: ω0 =ω2, ω1	2	-1359.91	1.98	0.177	0.496	$=\omega_0$
E. Two ratio: ω0 =ω1, ω2	2	-1363.87	1.97	0.378	$=\omega_0$	0.352

p: number of paramers in the model; l: log likelihood value; k: estimates of transition/transversion rate ratio; ω = estimates of dN/dS values; ω_0, ω_1 and ω_2: estimates of dN/dS ratios of outgroups (*Alangium* and *Davidia*), *CorPI*-A, *CorPI*-B in *Cornus* under different models, respectively.

Table 3. Results of Likelihood Ratio Tests (LRTs) for significance in differential selection force (ω = dN/dS) between *CorPI* of *Cornus* and *PI*-like genes in outgroup genera *Alangium* and *Davidia*. By comparing different models (from Table 2), null hypotheses were tested based on various assumptions.

Null Hypothesis	Assumption Made	Models Compared	2 * Δ l
I. $\omega_0=(\omega_1=\omega_2)$	$\omega_1=\omega_2$	B and C	10.88**
II. $\omega_0=\omega_1$	$\omega_0=\omega_2$	B and D	7.94**
III. $\omega_0=\omega_1$	ω_2 free	A and E	12.30**
IV. $\omega_0=\omega_2$	$\omega_0=\omega_1$	B and E	0.02
V. $\omega_0=\omega_2$	ω_1 free	A and D	4.38*

Likelihood Ratio Statistics is 2 *Δl, two times difference in likelihood scores derived from two models under comparison. One star * indicates significant ($p<5\%$; $\chi 2=3.84$) to reject null hypothesis; two stars ** indicate extremely significant ($p<1\%$; $\chi 2=6.63$) to reject null hypothesis.

Table 4. Parameters estimated for the *PISTILLATA*-like genes in *Cornus* based on various codon-based substitution models implemented in PAML 3.15 to test selection.

Model	p[a]	lnL[b]	dN/dS	Estimates of Parameters[c]	Positively Selected Sites[d]
M0: one-ratio	1	-1363.88	0.375	ω=0.375	None
M1: neutral (K=2)	1	-1334.83	0.268	P_0=0.876	None
M2: selection (K=3)	3	-1321.76	0.440	P_0=0.864, P_1=0.122, (P_2=0.015), ω_2=10.720	**: I_4
M3: discrete (K=3)	5	-1320.24	0.521	P_0=0.900, P_1=0.085, (P_2=0.015), ω_0=0.227, ω_1=1.618, ω_2=12.121	**: M_{54}, I_4, I_6 *: I_{11}
M7: beta	2	-1342.07	0.318	P=0.513, q=1.095	None
M8: beta&ω	4	-1327.47	0.414	P_0=0.933, (P_1=0.067), P=2.399, q=7.133, ω=2.691	**: I_4, I_6 *: M_{54}, I_{11}

a: Number of parameters in the model.

b: Log-likelihood scores; also see Yang et al. (2000) for the definitions of parameters.

c: P_i denotes the proportion of site falling in site class ω_i.

d: Sites potentially under positive selection identified under model M2, M3 and M8 with a posterior probability *: >95% or **: >99%.

Table 5. Comparisons of models (from Table 4) applied to the Likelihood Ratio Tests (LRTs) for testing heterogeneous selection at amino acid sites in *CorPI* genes. Stars indicate models that allow positive selection among amino acid sites (M2, M3 and M8) and significantly fit data better at >99% level.

LRT	Degrees of freedom	$\chi 2$ Critical Value (5%)	$\chi 2$ Critical Value (1%)	$2*\Delta l$
M0 vs. M3	4	9.49	13.28	87.28**
M1 vs. M2	2	5.99	9.21	26.14**
M7 vs. M8	2	5.99	9.21	29.20**

(*: P>95%; **: P>99%)

Table 6. Amino acid variation at positive selection sites (ω>1) identified at > 95% (*) and > 99% (**) levels suggested by various codon-based substitution models (Table 4).

Position of positively selected sites	Amino acid variation	M2 model (selection)	M3 model (discrete)	M8 model (beta&ω)
M_{54}	I,M,L	No	Yes**	Yes*
I_4	A,G,T,S,L	Yes**	Yes**	Yes**
I_6	S,P,T,K,R,N	No	Yes**	Yes**
I_{11}	L,M,S	No	Yes*	Yes*

Supplementary material. Parameter estimation of selection pressure based on codon-based likelihood method implemented in PAML 3.15. Model allows variable ω among all branches.

branch	t	N	S	dN/dS	dN	dS	N*dN	S*dS
72..1	0.014	164.3	39.7	0.0001	0	0.0248	0	1
72..2	0	164.3	39.7	0.1524	0	0	0	0
72..73	0.194	164.3	39.7	0.1618	0.0322	0.1988	5.3	7.9
73..74	0.032	164.3	39.7	0.0001	0	0.0554	0	2.2
74..75	0.018	164.3	39.7	0.0001	0	0.0315	0	1.2
75..76	0.015	164.3	39.7	981.667	0.0061	0	1	0
76..77	0.009	164.3	39.7	0.0001	0	0.016	0	0.6
77..78	0.104	164.3	39.7	0.1833	0.0186	0.1012	3	4
78..79	0	164.3	39.7	0.3972	0	0	0	0
79..3	0.023	164.3	39.7	0.0001	0	0.0387	0	1.5
79..80	0	164.3	39.7	0.3978	0	0	0	0
80..4	0.061	164.3	39.7	999	0.0251	0	4.1	0
80..81	0	164.3	39.7	0.3968	0	0	0	0
81..5	0	164.3	39.7	0.1988	0	0	0	0
81..82	0	164.3	39.7	0.3986	0	0	0	0
82..83	0	164.3	39.7	0.3983	0	0	0	0
83..6	0.017	164.3	39.7	970.7998	0.0068	0	1.1	0
83..84	0	164.3	39.7	0.1186	0	0	0	0
84..7	0.018	164.3	39.7	988.0298	0.0074	0	1.2	0
84..8	0.017	164.3	39.7	796.1066	0.0069	0	1.1	0
82..85	0	164.3	39.7	0.3977	0	0	0	0
85..9	0.015	164.3	39.7	976.3002	0.0061	0	1	0
85..86	0	164.3	39.7	0.3983	0	0	0	0
86..10	0	164.3	39.7	0.1989	0	0	0	0
86..11	0.015	164.3	39.7	971.5299	0.0061	0	1	0
78..87	0.031	164.3	39.7	0.2168	0.0061	0.0283	1	1.1
87..88	0.016	164.3	39.7	0.0001	0	0.028	0	1.1
88..89	0.015	164.3	39.7	974.6121	0.0061	0	1	0
89..12	0	164.3	39.7	0.2096	0	0	0	0
89..13	0.031	164.3	39.7	0.2181	0.0061	0.028	1	1.1
88..14	0.031	164.3	39.7	0.2169	0.0061	0.0279	1	1.1
87..90	0	164.3	39.7	0.3981	0	0	0	0
90..15	0.031	164.3	39.7	999	0.0128	0	2.1	0
90..16	0.096	164.3	39.7	1.1031	0.0326	0.0295	5.4	1.2
90..17	0.033	164.3	39.7	0.0001	0	0.0572	0	2.3
90..18	0.031	164.3	39.7	0.2155	0.0061	0.0282	1	1.1
90..91	0.03	164.3	39.7	999	0.0124	0	2	0

91..19	0.016	164.3	39.7	972.5335	0.0068	0	1.1	0
91..92	0.061	164.3	39.7	0.6726	0.0187	0.0277	3.1	1.1
92..20	0	164.3	39.7	0.1711	0	0	0	0
92..93	0.016	164.3	39.7	0.0001	0	0.0274	0	1.1
93..21	0.015	164.3	39.7	982.6769	0.0061	0	1	0
93..22	0.015	164.3	39.7	983.8859	0.0061	0	1	0
77..94	0.087	164.3	39.7	0.1219	0.0121	0.0993	2	3.9
94..95	0	164.3	39.7	0.3975	0	0	0	0
95..23	0	164.3	39.7	0.2259	0	0	0	0
95..24	0.015	164.3	39.7	984.7968	0.006	0	1	0
94..96	0.015	164.3	39.7	984.2976	0.0061	0	1	0
96..25	0	164.3	39.7	0.2137	0	0	0	0
96..26	0	164.3	39.7	0.2137	0	0	0	0
94..97	0.015	164.3	39.7	984.9729	0.006	0	1	0
97..27	0.015	164.3	39.7	0.0001	0	0.0264	0	1
97..28	0.015	164.3	39.7	994.5266	0.0061	0	1	0
94..98	0	164.3	39.7	0.3974	0	0	0	0
98..99	0	164.3	39.7	0.3984	0	0	0	0
99..29	0.06	164.3	39.7	0.6785	0.0185	0.0272	3	1.1
99..30	0.044	164.3	39.7	999	0.0183	0	3	0
98..100	0	164.3	39.7	0.3997	0	0	0	0
100..31	0.037	164.3	39.7	999	0.0154	0	2.5	0
100..32	0.029	164.3	39.7	999	0.0122	0	2	0
76..101	0.017	164.3	39.7	0.0001	0	0.0286	0	1.1
101..102	0.016	164.3	39.7	0.0001	0	0.028	0	1.1
102..33	0.015	164.3	39.7	990.9876	0.0061	0	1	0
102..34	0	164.3	39.7	0.2863	0	0	0	0
101..103	0.03	164.3	39.7	0.2346	0.0061	0.0261	1	1
103..35	0.061	164.3	39.7	0.702	0.0187	0.0266	3.1	1.1
103..104	0	164.3	39.7	0.3978	0	0	0	0
104..36	0	164.3	39.7	0.3234	0	0	0	0
104..37	0.015	164.3	39.7	997.2656	0.0062	0	1	0
75..105	0	164.3	39.7	0.3977	0	0	0	0
105..106	0.016	164.3	39.7	0.0001	0	0.0267	0	1.1
106..107	0.015	164.3	39.7	974.3642	0.0061	0	1	0
107..108	0	164.3	39.7	0.3979	0	0	0	0
108..109	0	164.3	39.7	0.4001	0	0	0	0
109..38	0.038	164.3	39.7	999	0.0159	0	2.6	0
109..39	0	164.3	39.7	0.2632	0	0	0	0
108..110	0.046	164.3	39.7	0.4636	0.0124	0.0268	2	1.1
110..40	0.016	164.3	39.7	0.0001	0	0.0269	0	1.1
110..111	0	164.3	39.7	0.3976	0	0	0	0
111..41	0	164.3	39.7	0.2785	0	0	0	0

111..42	0.015	164.3	39.7	989.195	0.0061	0	1	0
107..112	0	164.3	39.7	0.3969	0	0	0	0
112..113	0.015	164.3	39.7	996.4587	0.0062	0	1	0
113..43	0.063	164.3	39.7	0.1079	0.0081	0.0747	1.3	3
113..114	0.015	164.3	39.7	997.1863	0.0062	0	1	0
114..44	0.015	164.3	39.7	997.0345	0.0062	0	1	0
114..115	0	164.3	39.7	0.3982	0	0	0	0
115..45	0.015	164.3	39.7	999	0.0062	0	1	0
115..46	0.015	164.3	39.7	999	0.0062	0	1	0
112..116	0.03	164.3	39.7	0.2336	0.0062	0.0265	1	1.1
116..47	0.031	164.3	39.7	0.2343	0.0062	0.0266	1	1.1
116..117	0	164.3	39.7	0.3976	0	0	0	0
117..48	0.017	164.3	39.7	997.116	0.007	0	1.1	0
117..49	0.015	164.3	39.7	998.8038	0.0062	0	1	0
106..118	0.031	164.3	39.7	0.0001	0	0.0529	0	2.1
118..119	0	164.3	39.7	0.3947	0	0	0	0
119..120	0.03	164.3	39.7	0.2357	0.0062	0.0263	1	1
120..50	0.045	164.3	39.7	0.4802	0.0125	0.0261	2.1	1
120..51	0	164.3	39.7	0.3553	0	0	0	0
119..121	0.015	164.3	39.7	999	0.0062	0	1	0
121..52	0.03	164.3	39.7	999	0.0125	0	2.1	0
121..53	0.015	164.3	39.7	999	0.0062	0	1	0
118..122	0	164.3	39.7	0.3976	0	0	0	0
122..123	0.03	164.3	39.7	0.2386	0.0062	0.026	1	1
123..54	0	164.3	39.7	2.4757	0	0	0	0
123..55	0.019	164.3	39.7	975.2194	0.0079	0	1.3	0
122..124	0.015	164.3	39.7	0.0001	0	0.0261	0	1
124..56	0.015	164.3	39.7	999	0.0062	0	1	0
124..57	0	164.3	39.7	0.3876	0	0	0	0
105..125	0.03	164.3	39.7	999	0.0122	0	2	0
125..126	0	164.3	39.7	0.3979	0	0	0	0
126..58	0.015	164.3	39.7	990.4717	0.0061	0	1	0
126..59	0	164.3	39.7	0.2829	0	0	0	0
125..127	0	164.3	39.7	0.4049	0	0	0	0
127..60	0.016	164.3	39.7	984.9189	0.0065	0	1.1	0
127..61	0.031	164.3	39.7	0.2291	0.0062	0.0269	1	1.1
74..128	0.077	164.3	39.7	0.8306	0.0248	0.0298	4.1	1.2
128..129	0.049	164.3	39.7	0.3815	0.0125	0.0328	2.1	1.3
129..62	0	164.3	39.7	0.2206	0	0	0	0
129..130	0	164.3	39.7	0.3972	0	0	0	0
130..63	0	164.3	39.7	0.2216	0	0	0	0
130..131	0	164.3	39.7	0.3977	0	0	0	0
131..64	0.017	164.3	39.7	0.0001	0	0.0283	0	1.1

131..65	0.03	164.3	39.7	999	0.0124	0	2	0
128..132	0.098	164.3	39.7	0.0426	0.0061	0.1434	1	5.7
132..66	0.02	164.3	39.7	974.3669	0.0081	0	1.3	0
132..67	0.015	164.3	39.7	983.9236	0.0062	0	1	0
132..133	0.015	164.3	39.7	982.763	0.0061	0	1	0
133..68	0.019	164.3	39.7	955.3859	0.0079	0	1.3	0
133..69	0	164.3	39.7	0.2701	0	0	0	0
73..134	0.174	164.3	39.7	0.0514	0.0126	0.245	2.1	9.7
134..70	0.031	164.3	39.7	0.236	0.0063	0.0268	1	1.1
134..71	0	164.3	39.7	0.0001	0	0	0	0

t: branch length; N: number of nonsynonymous substitution sites; S: number of synonymous substitution sites; dN: nonsynonymous substitution rate; dS: synonymous substitution rate; dN/dS: ratio of nonsynonymous substitution rate versus synonymous substitution rate; N*dN: number of nonsynonymous substitution changes; S*dS: number of synonymous substitution changes.

Supplementary material. Estimations of divergence times and nucleotide substitution rates for each node and each branch of the phylogeny (Fig 9) based on *CorPI*-A coding and noncoding regions using r8s 1.71.

Nodes	Age	ESR	Nodes	Age	ESR	Nodes	Age	ESR
1	90.04	-	C.disc02 08 23	0	2.22E-04	45	40.42	8.79E-04
Fixed	62	1.97E-03	24	21.49	1.61E-03	46	8.74	9.03E-04
3	48.29	3.36E-03	C.flor01 148 37	0	3.55E-04	C.cont02 94 4	0	1.05E-03
4	27.4	2.98E-03	26	16.92	1.89E-03	48	6.48	5.23E-04
5	12.21	2.13E-03	C.flor01 148 7	0	4.25E-04	C.cont02 94 15	0	2.05E-04
6	10.73	1.65E-03	C.flor01 148 5	0	3.16E-03	C.cont02 94 53	0	5.09E-04
7	3.74	1.53E-03	29	3.07	4.09E-03	51	27.13	6.93E-04
8	3.13	1.62E-03	30	0.98	4.01E-03	52	7.92	5.31E-04
C.kous04 C45 1	0	1.60E-03	31	0.8	4.13E-03	C.also02 206 38	0	6.52E-04
C.kous04 C45 27	0	1.72E-03	C.cana04 01 7	0	4.23E-03	C.also02 206 26	0	1.66E-04
C.kous04 C45 19	0	1.38E-03	C.cana04 01 23	0	4.12E-03	C.alba02 269 5	0	6.47E-04
12	10.29	1.35E-03	C.cana04 01 35	0	3.86E-03	56	27.51	7.02E-04
C.angu02 46 16	0	1.18E-03	C.suec27 2 69	0	4.21E-03	C.also02 206 5	0	7.12E-04
C.angu02 46 71	0	1.22E-03	36	39.01	8.84E-04	58	12.4	4.42E-04
C.angu02 46 23	0	2.11E-03	37	3.92	3.90E-04	C.alba02 269 1	0	4.78E-04
16	25.59	2.12E-03	C.chin02 83 9	0	4.10E-04	C.alba02 269 3	0	2.34E-04
17	17.15	1.58E-03	C.chin02 83 12	0	3.03E-04	61	4.94	1.32E-03
18	12.73	8.26E-04	40	6.28	5.13E-04	62	1.75	1.71E-03
C.disc02 08 4	0	6.34E-04	C.offi918 85A 29	0	2.98E-04	C.oblo02 223 20	0	1.87E-03
C.disc02 08 41	0	3.91E-04	C.offi918 85A 33	0	5.83E-04	C.oblo02 223 31	0	1.67E-03
21	6.92	1.44E-03	43	52.52	1.39E-03	C.oblo02 223 27	0	8.46E-04
C.disc02 08 17	0	1.90E-03	44	44.02	1.01E-03	C.oblo02 223 32	0	1.07E-03

Unit for divergence time estimation is million years before present (MYBP), and for estimation of substitution rates (ESR) is substitutions/site/million year. Node before diversification of CC and (BB+DW) was fixed at 62 mya based on fossils. Program cannot estimate ESR for node 1, the basal branch. Node numbers correspond to internodes indicated in tree (Fig 9); sample names indicate terminal branches.

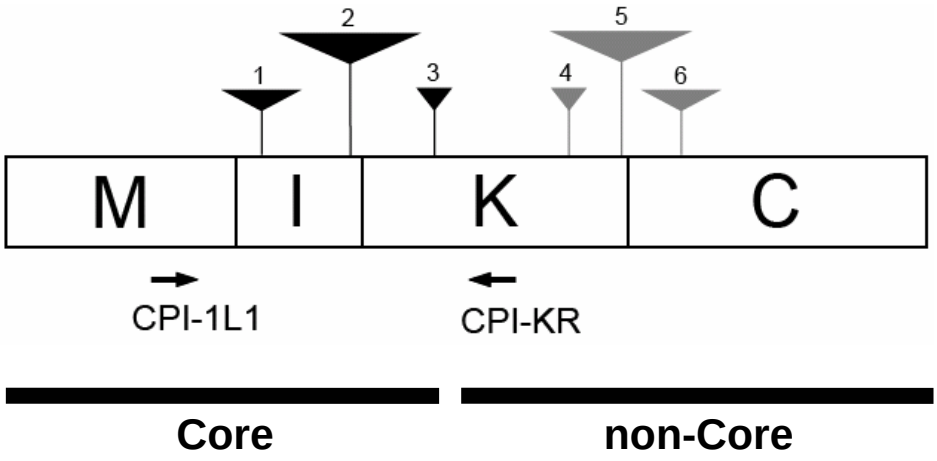
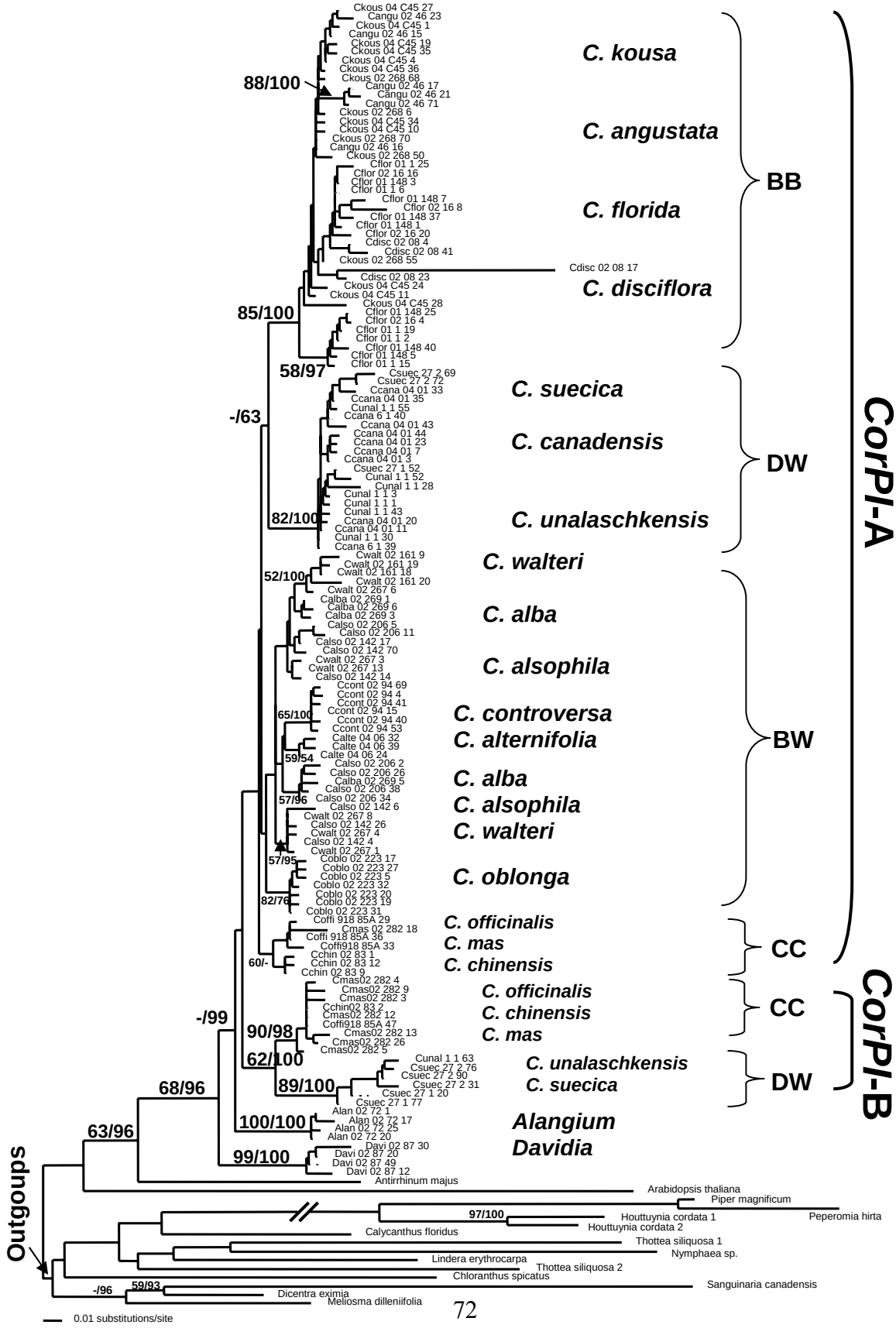
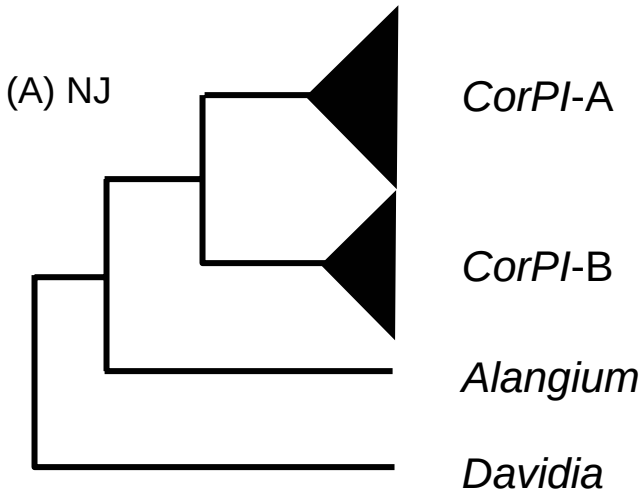
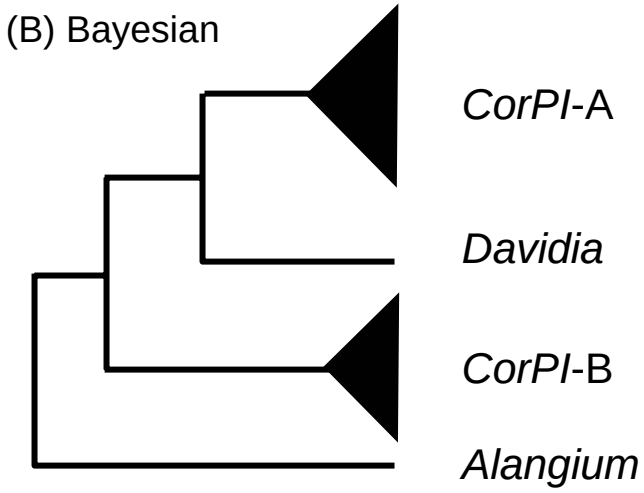


Figure 1. Schematic diagram of MIKC structure of *PISTILLATA* homologs based on complete sequence of *Antirrhinum major* (X68831). The region between primers CPI-1L1 and CPI-KR is analyzed for this study. Core and non-core regions based on protein activity study are indicated. Reversed triangles indicate position and relative sizes of six introns. Introns 1, 2 and 3 are covered in this study.

Figure 2. Genealogy of coding sequences of *PI*-like genes inferred from NJ analysis rooted by basal eudicots and basal angiosperms. Numbers near branches are NJ bootstrap values/Bayesian posterior probabilities. Only nodes supported by ⩾ 50% bootstrap value and ⩾ 60% Bayesian posterior probability are indicated. A & B: two schematic alternative topologies differing in placement of *Davidia* resulting from NJ and Bayesian analyses, respectively. Abbreviations of species names indicated in Table 1. DW = dwarf dogwoods, CC = cornelian cherry, BW = blue- or white-fruited dogwoods, and BB = big-bracted dogwoods. *CorPI*-A and *CorPI*-B indicate two ancient copies of *CorPI* gene clades found in *Cornus*.


Ckous 04 C45 27
Cangu 02 46 23
Ckous 04 C45 1
Cangu 02 46 15
Ckous 04 C45 19
Ckous 04 C45 35
Ckous 04 C45 4
Ckous 04 C45 36
Ckous 02 268 68
Cangu 02 46 17
Cangu 02 46 21
Cangu 02 46 71
Ckous 02 268 6
Ckous 04 C45 34
Ckous 04 C45 10
Ckous 02 268 70
Cangu 02 46 16
Ckous 02 268 50
Cflor 01 1 25
Cflor 02 16 16
Cflor 01 148 3
Cflor 01 1 6
Cflor 01 148 7
Cflor 02 16 8
Cflor 01 148 37
Cflor 01 148 1
Cflor 02 16 20
Cdisc 02 08 4
Cdisc 02 08 41
Ckous 02 268 55
Cdisc 02 08 17
Cdisc 02 08 23
Ckous 04 C45 24
Ckous 04 C45 11
Ckous 04 C45 28
Cflor 01 148 25
Cflor 02 16 4
Cflor 01 1 19
Cflor 01 1 2
Cflor 01 148 40
Cflor 01 148 5
Cflor 01 1 15
Csuec 27 2 69
Csuec 27 2 72
Ccana 04 01 33
Ccana 04 01 35
Cunal 1 1 55
Ccana 6 1 40
Ccana 04 01 43
Ccana 04 01 44
Ccana 04 01 23
Ccana 04 01 7
Ccana 04 01 3
Csuec 27 1 52
Cunal 1 1 52
Cunal 1 1 28
Cunal 1 1 3
Cunal 1 1 1
Cunal 1 1 43
Ccana 04 01 20
Ccana 04 01 11
Cunal 1 1 30
Ccana 6 1 39
Cwalt 02 161 9
Cwalt 02 161 19
Cwalt 02 161 18
Cwalt 02 161 20
Cwalt 02 267 6
Calba 02 269 1
Calba 02 269 6
Calba 02 269 3
Calso 02 206 5
Calso 02 206 11
Calso 02 142 17
Calso 02 142 70
Cwalt 02 267 3
Cwalt 02 267 13
Calso 02 142 14
Ccont 02 94 69
Ccont 02 94 4
Ccont 02 94 41
Ccont 02 94 15
Ccont 02 94 40
Ccont 02 94 53
Calte 04 06 32
Calte 04 06 39
Calte 04 06 24
Calso 02 206 2
Calso 02 206 26
Calba 02 269 5
Calso 02 206 38
Calso 02 206 34
Calso 02 142 6
Cwalt 02 267 8
Calso 02 142 26
Cwalt 02 267 4
Calso 02 142 4
Cwalt 02 267 1
Coblo 02 223 17
Coblo 02 223 27
Coblo 02 223 5
Coblo 02 223 32
Coblo 02 223 20
Coblo 02 223 19
Coblo 02 223 31
Coffi 918 85A 29
Cmas 02 282 18
Coffi 918 85A 36
Coffi918 85A 33
Cchin 02 83 1
Cchin 02 83 12
Cchin 02 83 9
Cmas02 282 4
Cmas02 282 9
Cmas02 282 3
Cchin02 83 2
Cmas02 282 12
Coffi918 85A 47
Cmas02 282 13
Cmas02 282 26
Cmas02 282 5
Cunal 1 1 63
Csuec 27 2 76
Csuec 27 2 90
Csuec 27 2 31
Csuec 27 1 20
Csuec 27 1 77
Alan 02 72 1
Alan 02 72 17
Alan 02 72 25
Alan 02 72 20
Davi 02 87 30
Davi 02 87 20
Davi 02 87 49
Davi 02 87 12
Antirrhinum majus
Arabidopsis thaliana
Piper magnificum
Houttuynia cordata 1
Houttuynia cordata 2
Peperomia hirta
Calycanthus floridus
Thottea siliquosa 1
Nymphaea sp.
Lindera erythrocarpa
Thottea siliquosa 2
Chloranthus spicatus
Sanguinaria canadensis
Dicentra eximia
Meliosma dilleniifolia
0.01 substitutions/site
88/100
85/100
58/97
-/63
82/100
52/100
65/100
59/54
57/96
57/95
82/76
60/-
-/99
90/98
62/100
89/100
68/96
100/100
63/96
99/100
97/100
-/96
59/93
C. kousa
C. angustata
C. florida
C. disciflora
C. suecica
C. canadensis
C. unalaschkensis
C. walteri
C. alba
C. alsophila
C. controversa
C. alternifolia
C. alba
C. alsophila
C. walteri
C. oblonga
C. officinalis
C. mas
C. chinensis
C. officinalis
C. chinensis
C. mas
C. unalaschkensis
C. suecica
Alangium
Davidia
BB
DW
BW
CC
CC
DW
CorPI-A
CorPI-B
Outgroups


(A) NJ
CorPI-A
CorPI-B
Alangium
Davidia


(B) Bayesian
CorPI-A
Davidia
CorPI-B
Alangium

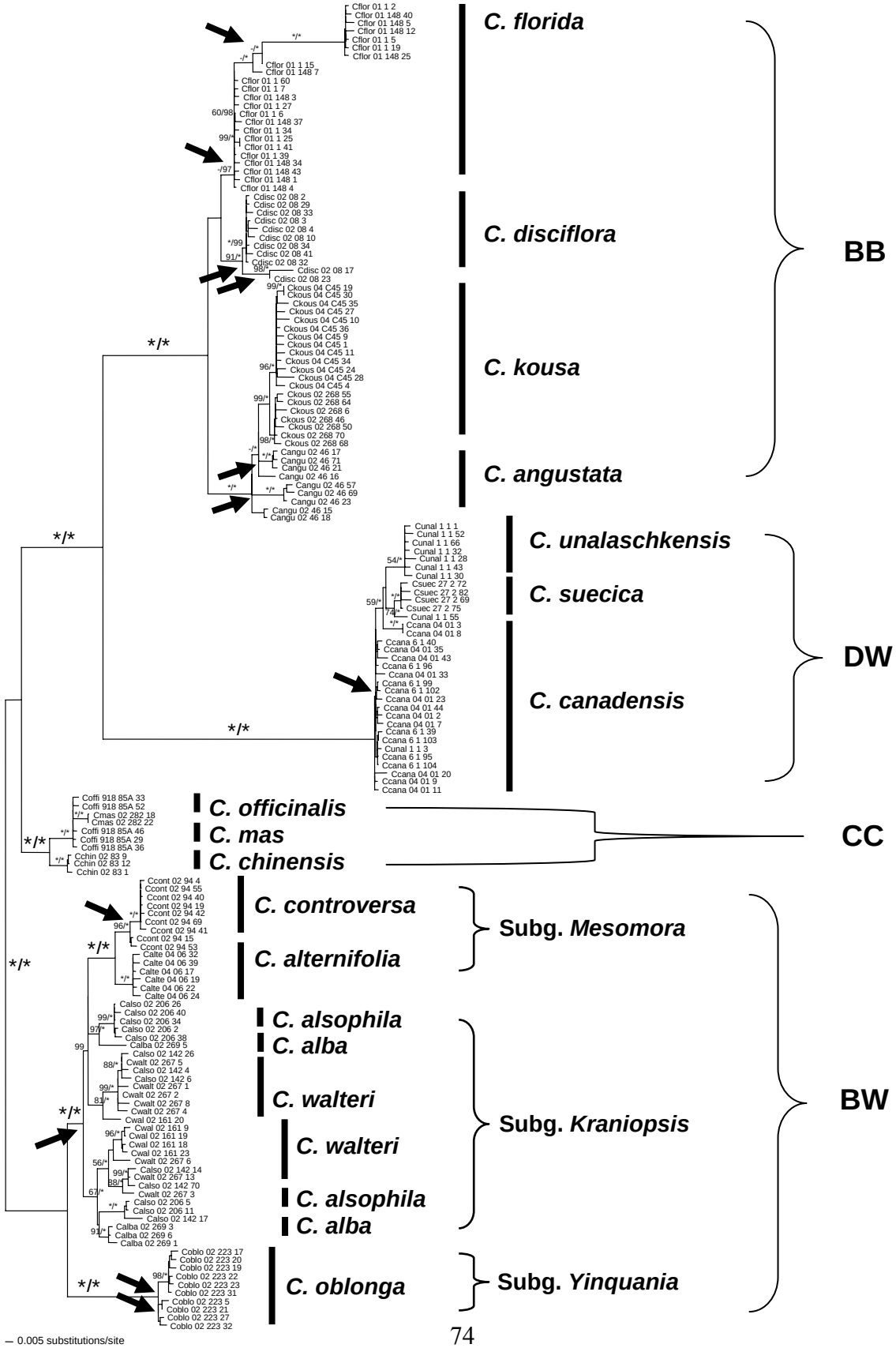

C. florida
C. disciflora
C. kousa
C. angustata
BB
C. unalaschkensis
C. suecica
C. canadensis
DW
C. officinalis
C. mas
C. chinensis
CC
C. controversa
Subg. Mesomora
C. alternifolia
C. alsophila
C. alba
C. walteri
C. walteri
C. alsophila
C. alba
Subg. Kraniopsis
BW
C. oblonga
Subg. Yinquania
— 0.005 substitutions/site

Figure 3. Phylogram of one *CorPI*-A gene tree based on Bayesian analyses including both intron and exon regions without rooting. Numbers next to nodes are Neighbor-Joining bootstraps/Bayesian posterior probabilities. Star and hyphen indicate support of 100% and ≤ 50%, respectively. Arrows indicate major duplication events identified based on both phylogeny and intron length differences (also see Fig 4). Abbreviations of taxon names are explained in Table 1. DW = dwarf dogwoods, CC = cornelian cherry, BW = blue- or white-fruited dogwoods, and BB = big-bracted dogwoods. Subgenera within BW subgroups are also indicated.

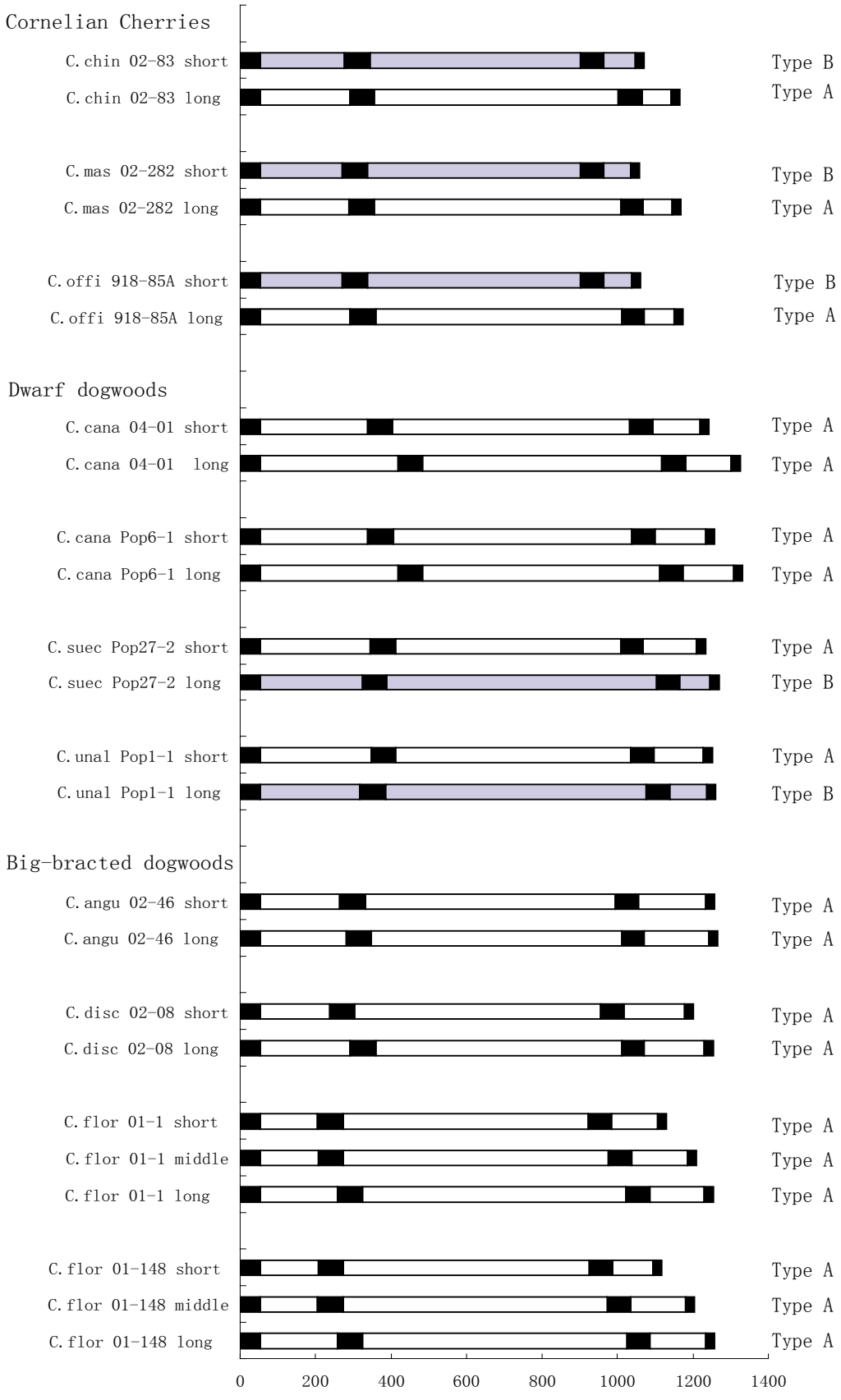

Cornelian Cherries
C.chin 02-83 short
Type B
C.chin 02-83 long
Type A
C.mas 02-282 short
Type B
C.mas 02-282 long
Type A
C.offi 918-85A short
Type B
C.offi 918-85A long
Type A
Dwarf dogwoods
C.cana 04-01 short
Type A
C.cana 04-01 long
Type A
C.cana Pop6-1 short
Type A
C.cana Pop6-1 long
Type A
C.suec Pop27-2 short
Type A
C.suec Pop27-2 long
Type B
C.unal Pop1-1 short
Type A
C.unal Pop1-1 long
Type B
Big-bracted dogwoods
C.angu 02-46 short
Type A
C.angu 02-46 long
Type A
C.disc 02-08 short
Type A
C.disc 02-08 long
Type A
C.flor 01-1 short
Type A
C.flor 01-1 middle
Type A
C.flor 01-1 long
Type A
C.flor 01-148 short
Type A
C.flor 01-148 middle
Type A
C.flor 01-148 long
Type A
0
200
400
600
800
1000
1200
1400

Figure 4. Comparisons of introns among copies of *CorPI*-A and *CorPI*-B with length differences. Black boxes indicate exons, and open/gray boxes are introns. Intron regions of Type A and Type B copies are highly divergent, which are demonstrated by open and gray boxes respectively. Lengths of boxes are proportional to sequence size.

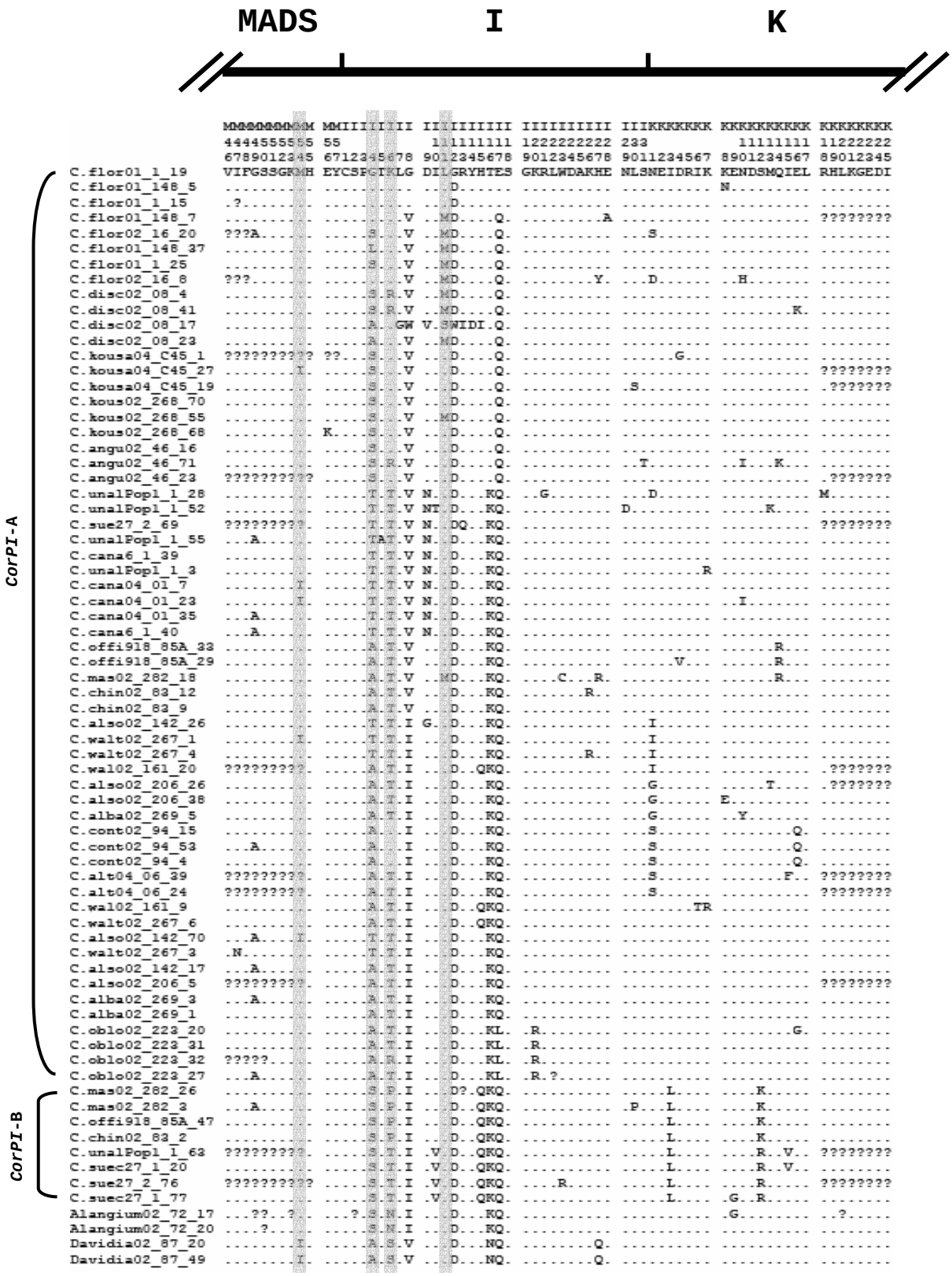
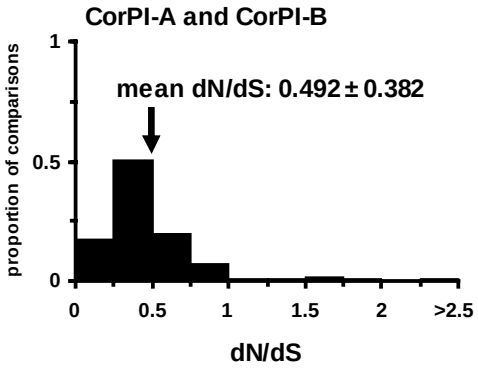
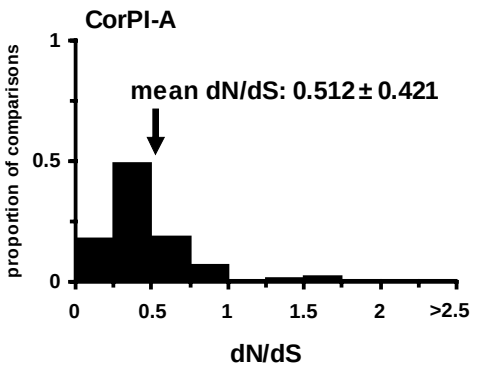
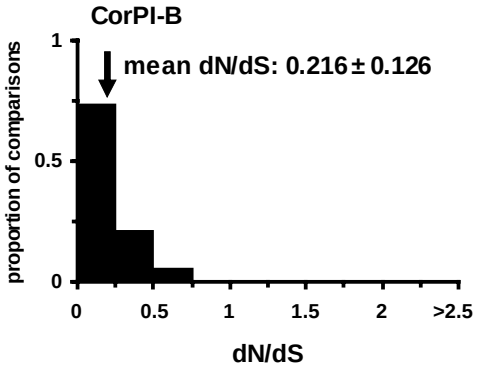
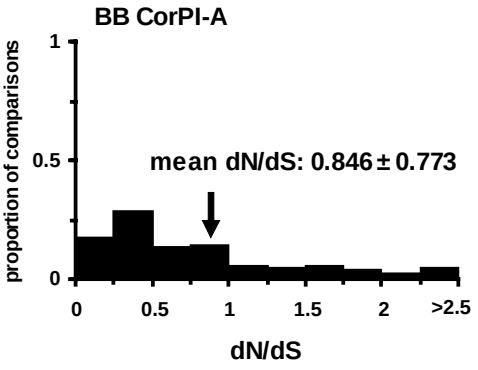

MADS
I
K
CorPI-A
CorPI-B

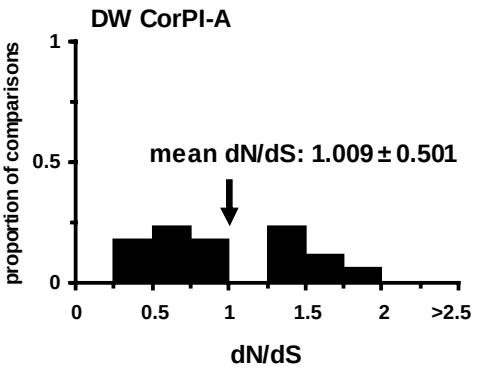
Figure 5. Alignment of coding sequences of 72 *PI*-like genes that cover whole I-domain and partial MADS- and K-domains from *Cornus* and its outgroup genera *Alangium* and *Davidia*. C. disc02-08-17 showing accelerated changes at beginning of I domain (also see results) is excluded from MEGA and PAML analyses. Positively selected sites (M_{54}, I_4, I_6, I_{11}) detected applying codon-based substitution models (see Table 4 and 6) are shaded in gray. Two major sequence types identified in *Cornus* are indicated. Amino acid labels, M: MADS domain aa residues, I: I domain aa residues, K: K domain aa residues.

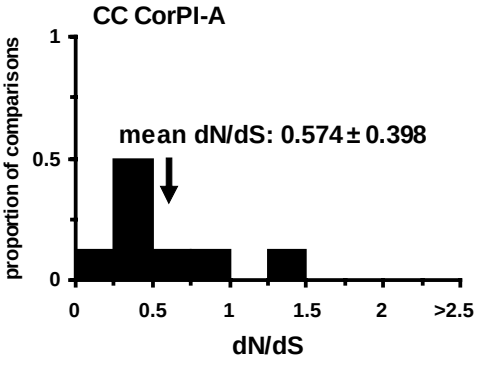

CorPI-A and CorPI-B
mean dN/dS: 0.492 ± 0.382
proportion of comparisons
1
0.5
0
0
0.5
1
1.5
2
>2.5
dN/dS


CorPI-A
mean dN/dS: 0.512 ± 0.421
proportion of comparisons
1
0.5
0
0
0.5
1
1.5
2
>2.5
dN/dS


CorPI-B
mean dN/dS: 0.216 ± 0.126
proportion of comparisons
1
0.5
0
0
0.5
1
1.5
2
>2.5
dN/dS


BB CorPI-A
mean dN/dS: 0.846 ± 0.773
proportion of comparisons
1
0.5
0
0
0.5
1
1.5
2
>2.5
dN/dS


DW CorPI-A
mean dN/dS: 1.009 ± 0.501
proportion of comparisons
1
0.5
0
0
0.5
1
1.5
2
>2.5
dN/dS


CC CorPI-A
mean dN/dS: 0.574 ± 0.398
proportion of comparisons
1
0.5
0
0
0.5
1
1.5
2
>2.5
dN/dS

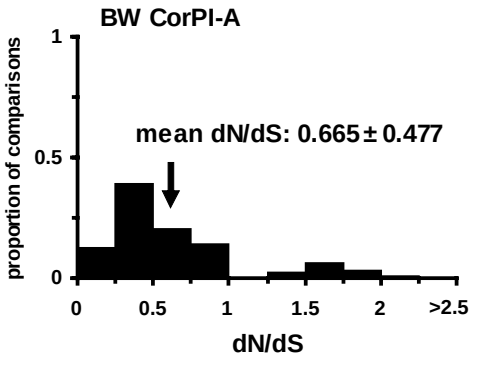
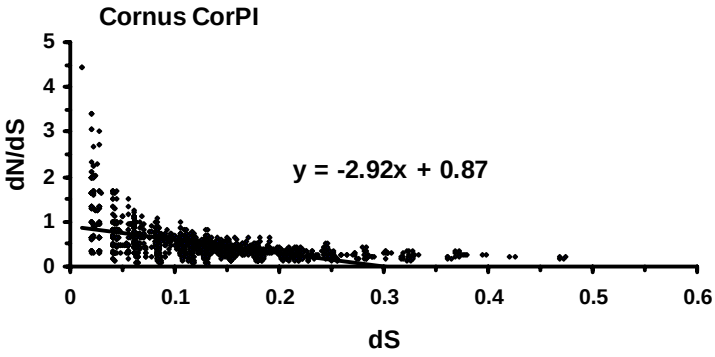
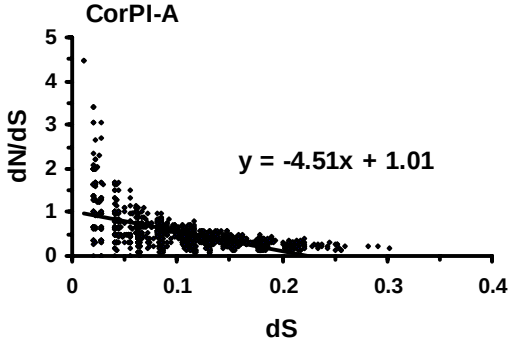
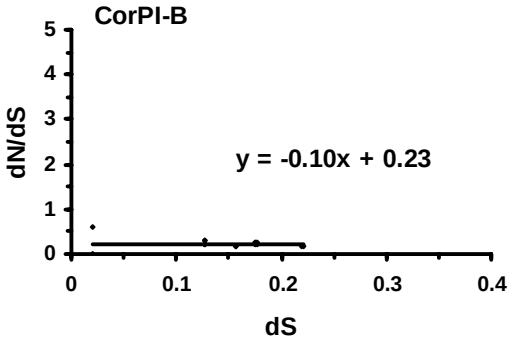
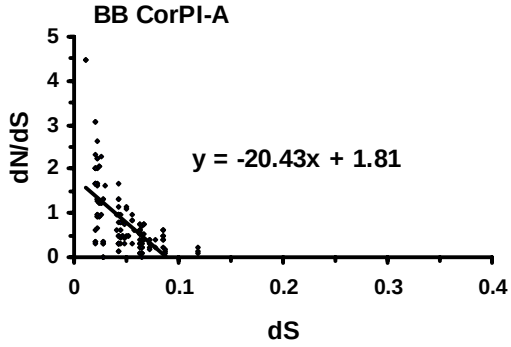

BW CorPI-A
mean dN/dS: 0.665 ± 0.477
proportion of comparisons
1
0.5
0
0
0.5
1
1.5
2
>2.5
dN/dS

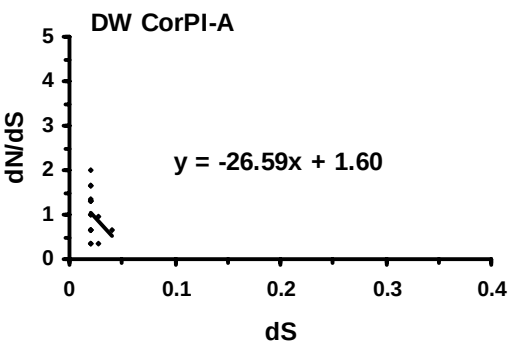
Figure 6. Distribution of dN/dS ratios based on pairwise comparisons (modified Nei-Gojobori, Jukes-Cantor model implemented in MEGA 3.1) for *CorPI* genes in different partitions for all *CorPI* genes, *CorPI*-A copy, *CorPI*-B copy, and for each of four morphological subgroups of *CorPI*-A copy, with mean values and standard deviations indicated. Pairwise comparisons that had dS = 0 are excluded from analyses.

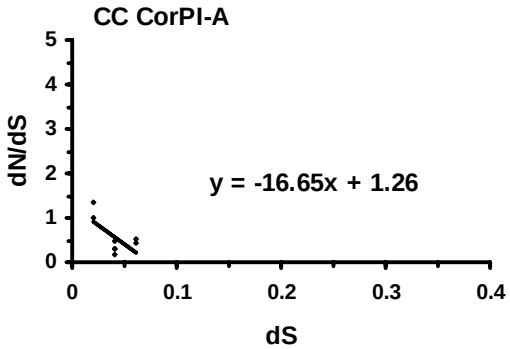

Cornus CorPI
dN/dS
y = -2.92x + 0.87
dS


CorPI-A
dN/dS
y = -4.51x + 1.01
dS


CorPI-B
dN/dS
y = -0.10x + 0.23
dS


BB CorPI-A
dN/dS
y = -20.43x + 1.81
dS


DW CorPI-A
dN/dS
y = -26.59x + 1.60
dS


CC CorPI-A
dN/dS
y = -16.65x + 1.26
dS

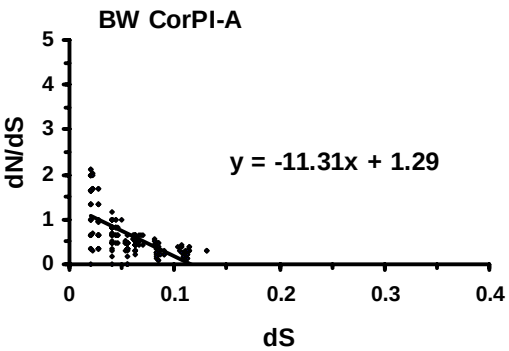

BW CorPI-A
dN/dS
y = -11.31x + 1.29
dS

Figure 7. Plots of dN/dS ratios versus dS based on estimates of pairwise comparisons (modified Nei-Gojobori, Jukes-Cantor model) using MEGA 3.1 for all *CorPI* genes, *CorPI*-A copy, *CorPI*-B copy, and for each of four morphological subgroups of *CorPI*-A copy. Regression lines and equations indicate trend of relationships between dN/dS and dS. Pairwise comparisons that had dS = 0 are excluded from analyses.

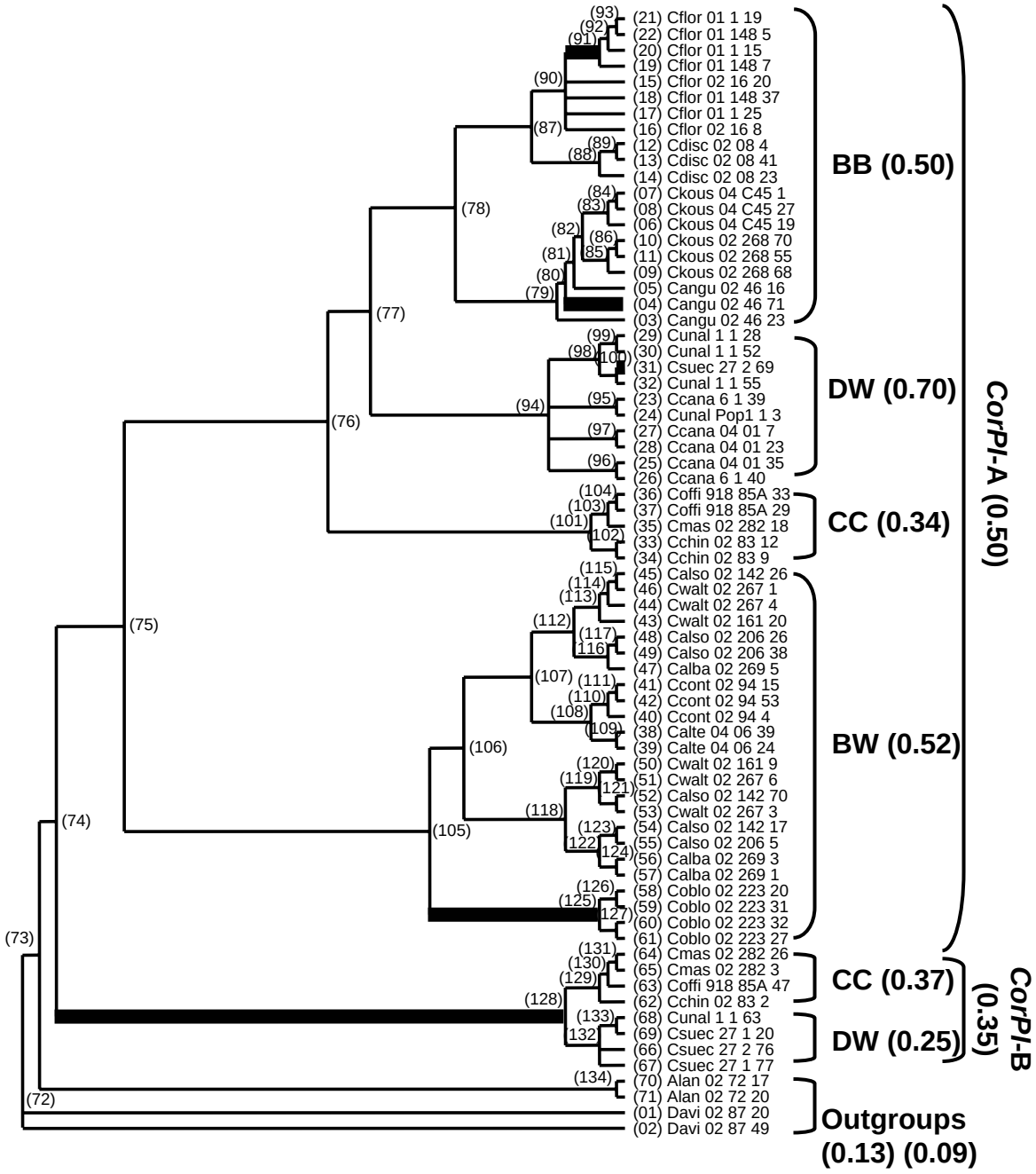

(21) Cflor 01 1 19
(22) Cflor 01 148 5
(20) Cflor 01 1 15
(19) Cflor 01 148 7
(15) Cflor 02 16 20
(18) Cflor 01 148 37
(17) Cflor 01 1 25
(16) Cflor 02 16 8
(12) Cdisc 02 08 4
(13) Cdisc 02 08 41
(14) Cdisc 02 08 23
(07) Ckous 04 C45 1
(08) Ckous 04 C45 27
(06) Ckous 04 C45 19
(10) Ckous 02 268 70
(11) Ckous 02 268 55
(09) Ckous 02 268 68
(05) Cangu 02 46 16
(04) Cangu 02 46 71
(03) Cangu 02 46 23
BB (0.50)
(29) Cunal 1 1 28
(30) Cunal 1 1 52
(31) Csuec 27 2 69
(32) Cunal 1 1 55
(23) Ccana 6 1 39
(24) Cunal Pop1 1 3
(27) Ccana 04 01 7
(28) Ccana 04 01 23
(25) Ccana 04 01 35
(26) Ccana 6 1 40
DW (0.70)
(36) Coffi 918 85A 33
(37) Coffi 918 85A 29
(35) Cmas 02 282 18
(33) Cchin 02 83 12
(34) Cchin 02 83 9
CC (0.34)
(45) Calso 02 142 26
(46) Cwalt 02 267 1
(44) Cwalt 02 267 4
(43) Cwalt 02 161 20
(48) Calso 02 206 26
(49) Calso 02 206 38
(47) Calba 02 269 5
(41) Ccont 02 94 15
(42) Ccont 02 94 53
(40) Ccont 02 94 4
(38) Calte 04 06 39
(39) Calte 04 06 24
(50) Cwalt 02 161 9
(51) Cwalt 02 267 6
(52) Calso 02 142 70
(53) Cwalt 02 267 3
(54) Calso 02 142 17
(55) Calso 02 206 5
(56) Calba 02 269 3
(57) Calba 02 269 1
(58) Coblo 02 223 20
(59) Coblo 02 223 31
(60) Coblo 02 223 32
(61) Coblo 02 223 27
BW (0.52)
CorPI-A (0.50)
(64) Cmas 02 282 26
(65) Cmas 02 282 3
(63) Coffi 918 85A 47
(62) Cchin 02 83 2
CC (0.37)
(68) Cunal 1 1 63
(69) Csuec 27 1 20
(66) Csuec 27 2 76
(67) Csuec 27 1 77
DW (0.25)
CorPI-B (0.35)
(70) Alan 02 72 17
(71) Alan 02 72 20
(01) Davi 02 87 20
(02) Davi 02 87 49
Outgroups (0.13) (0.09)
(72)
(73)
(74)
(75)
(76)
(77)
(78)
(79)
(80)
(81)
(82)
(83)
(84)
(85)
(86)
(87)
(88)
(89)
(90)
(91)
(92)
(93)
(94)
(95)
(96)
(97)
(98)
(99)
(100)
(101)
(102)
(103)
(104)
(105)
(106)
(107)
(108)
(109)
(110)
(111)
(112)
(113)
(114)
(115)
(116)
(117)
(118)
(119)
(120)
(121)
(122)
(123)
(124)
(125)
(126)
(127)
(128)
(129)
(130)
(131)
(132)
(133)
(134)

Figure 8. *CorPI* phylogeny compatible to coding and entire region analyses (Fig 2 and Fig 3) applied for PAML analyses. (1) To estimate the dN/dS values for two *CorPI* paralogs, *CorPI*-A is defined as branch 1 and *CorPI*-B as branch 2; background as undefined braches 0 for testing rate differences between two copies (also see Materials and Methods). (2) To estimate dN/dS values for different morphological subgroups, each subgroup is given a branch number that allow different dN/dS ratios among these lineages. Numbers in parentheses correspond to estimation of dN/dS values for each branch (see supplementary materials). Two dN/dS values of outgroups are based on two runs of analyses mentioned above. Thick branches indicate significant increases of dN/dS values detected with ≥2 nonsynonymous substitutions but 0 synonymous substitutions (see results and discussions). Abbreviations of taxa names are explained in Table 1. DW = dwarf dogwoods, CC = cornelian cherry, BW = blue- or white-fruited dogwoods, and BB = big-bracted dogwoods.

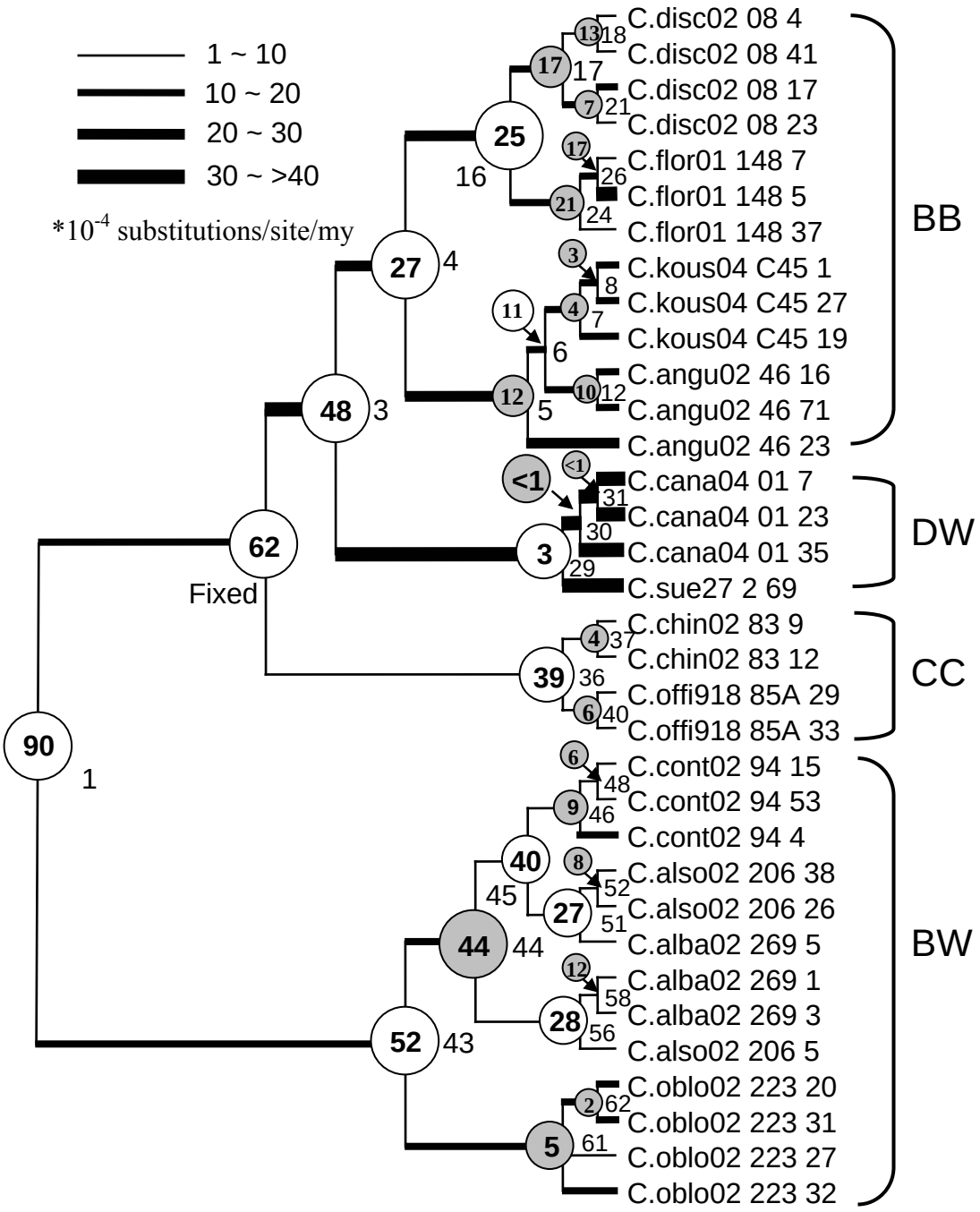
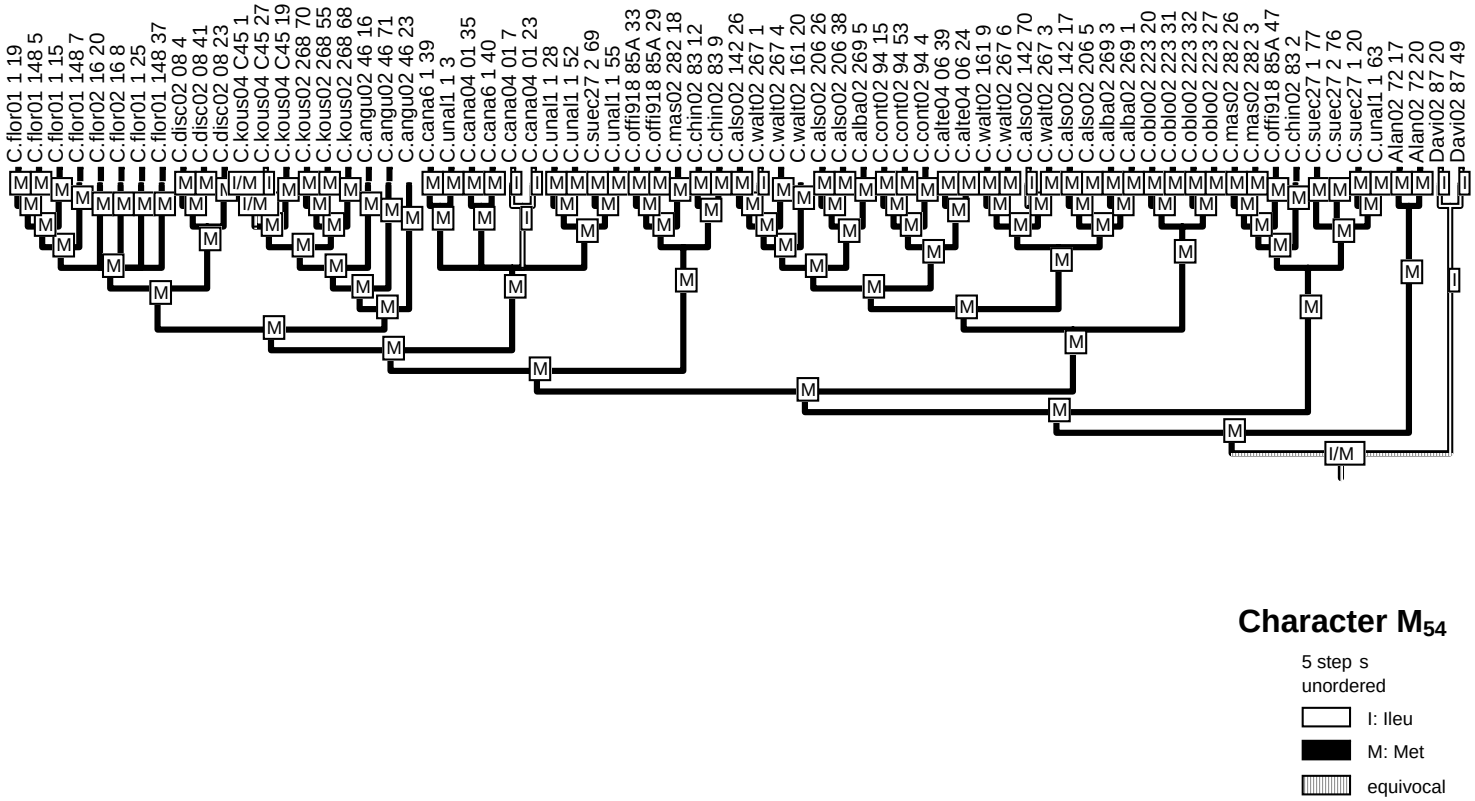
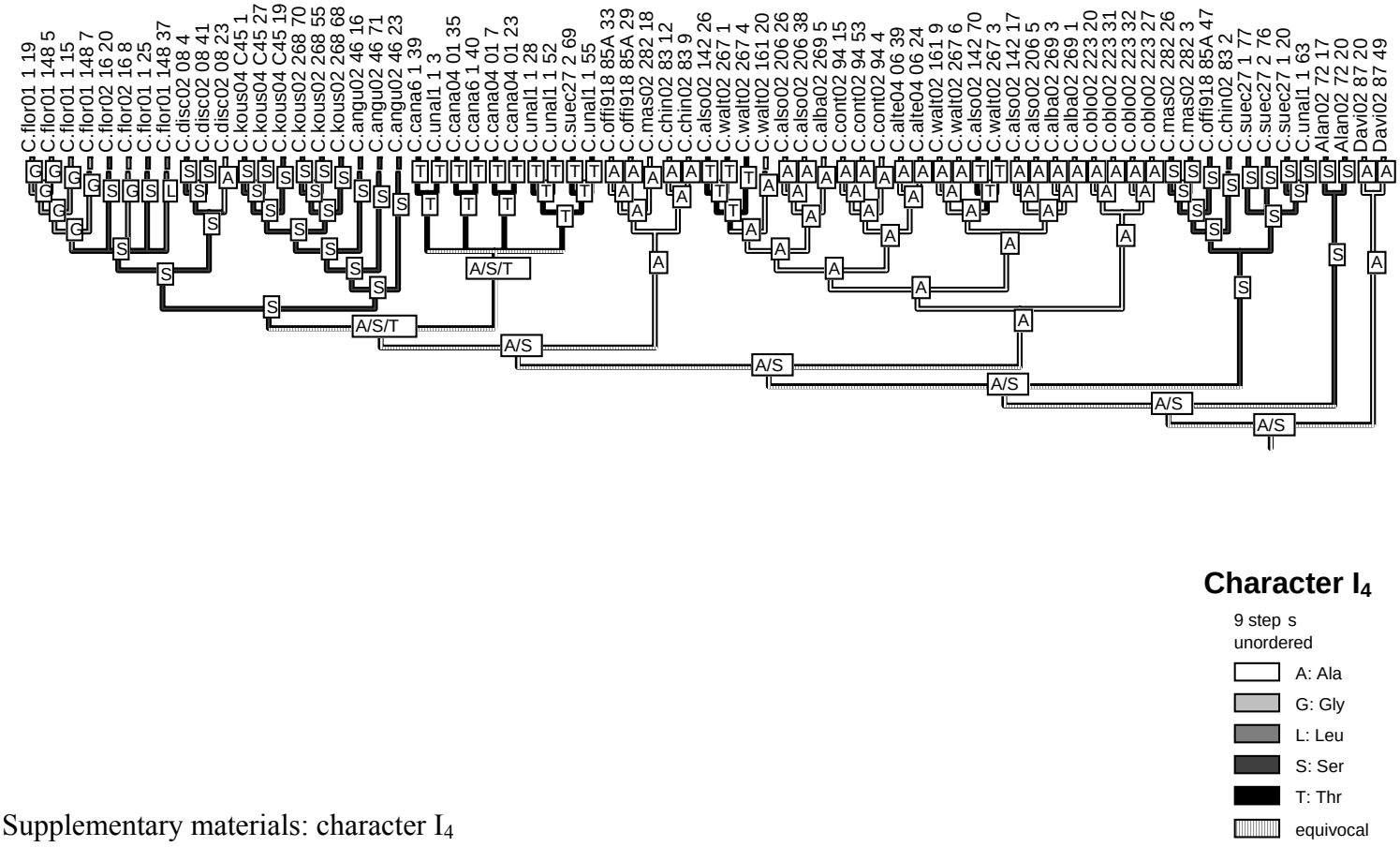

1 ~ 10
10 ~ 20
20 ~ 30
30 ~ >40
*10-4 substitutions/site/my
C.disc02 08 4
C.disc02 08 41
C.disc02 08 17
C.disc02 08 23
C.flor01 148 7
C.flor01 148 5
C.flor01 148 37
C.kous04 C45 1
C.kous04 C45 27
C.kous04 C45 19
C.angu02 46 16
C.angu02 46 71
C.angu02 46 23
C.cana04 01 7
C.cana04 01 23
C.cana04 01 35
C.sue27 2 69
C.chin02 83 9
C.chin02 83 12
C.offi918 85A 29
C.offi918 85A 33
C.cont02 94 15
C.cont02 94 53
C.cont02 94 4
C.also02 206 38
C.also02 206 26
C.alba02 269 5
C.alba02 269 1
C.alba02 269 3
C.also02 206 5
C.oblo02 223 20
C.oblo02 223 31
C.oblo02 223 27
C.oblo02 223 32
BB
DW
CC
BW
Fixed

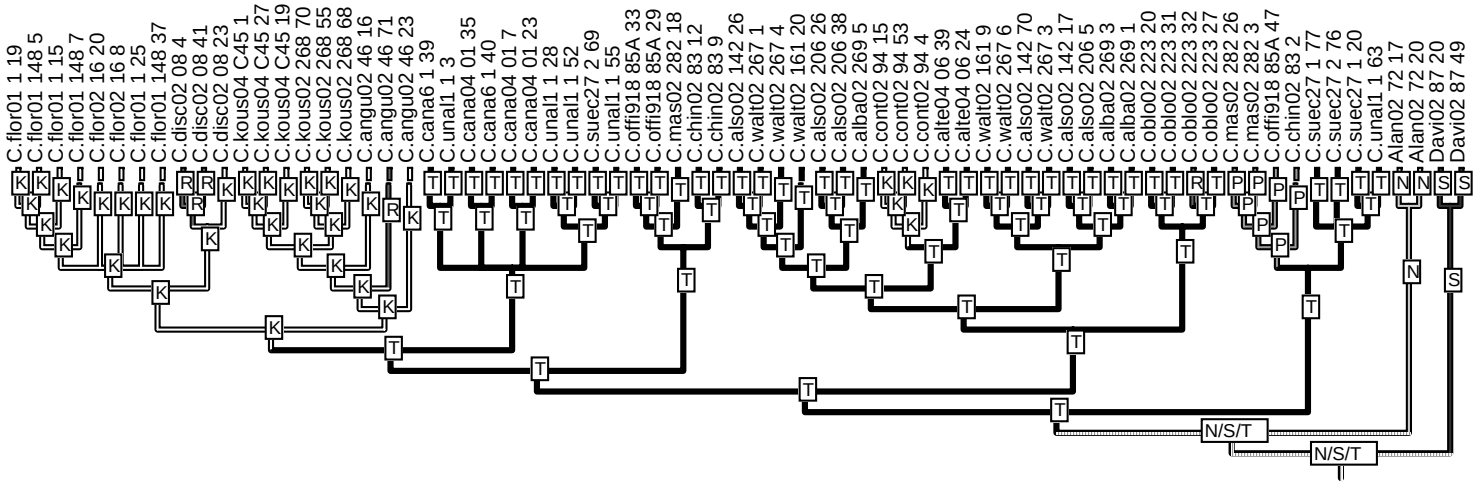
Figure 9. Estimation of divergence times and evolutionary rates of *CorPI*-A including both intron and exon regions in r8s 1.71. Topology is constrained to agree with all NJ and Bayesain analyses; branch lengths are estimated by Maximum Likelihood analysis. Node of diversification of CC and (BB+DW) is fixed at 62 mya based on fossils (the oldest fossil of CC group is from the late Paleocene; Crane et al. 1990; Xiang et al. 2003). Numbers in open circles indicate estimation of divergence time of speciation events; numbers in gray circles indicate the time of gene duplication events. Estimates of divergence time of this analysis should be taken as minimum times for those nodes. Thickness of branch indicates relative magnitude of substitution rates ($*10^{-4}$ substitutions/site/million year). DW = dwarf dogwoods, CC = cornelian cherry, BW = blue- or white-fruited dogwoods, and BB = big-bracted dogwoods. Numbers near internodes correspond to results of substitution rates for each branch reported in supplementary materials.



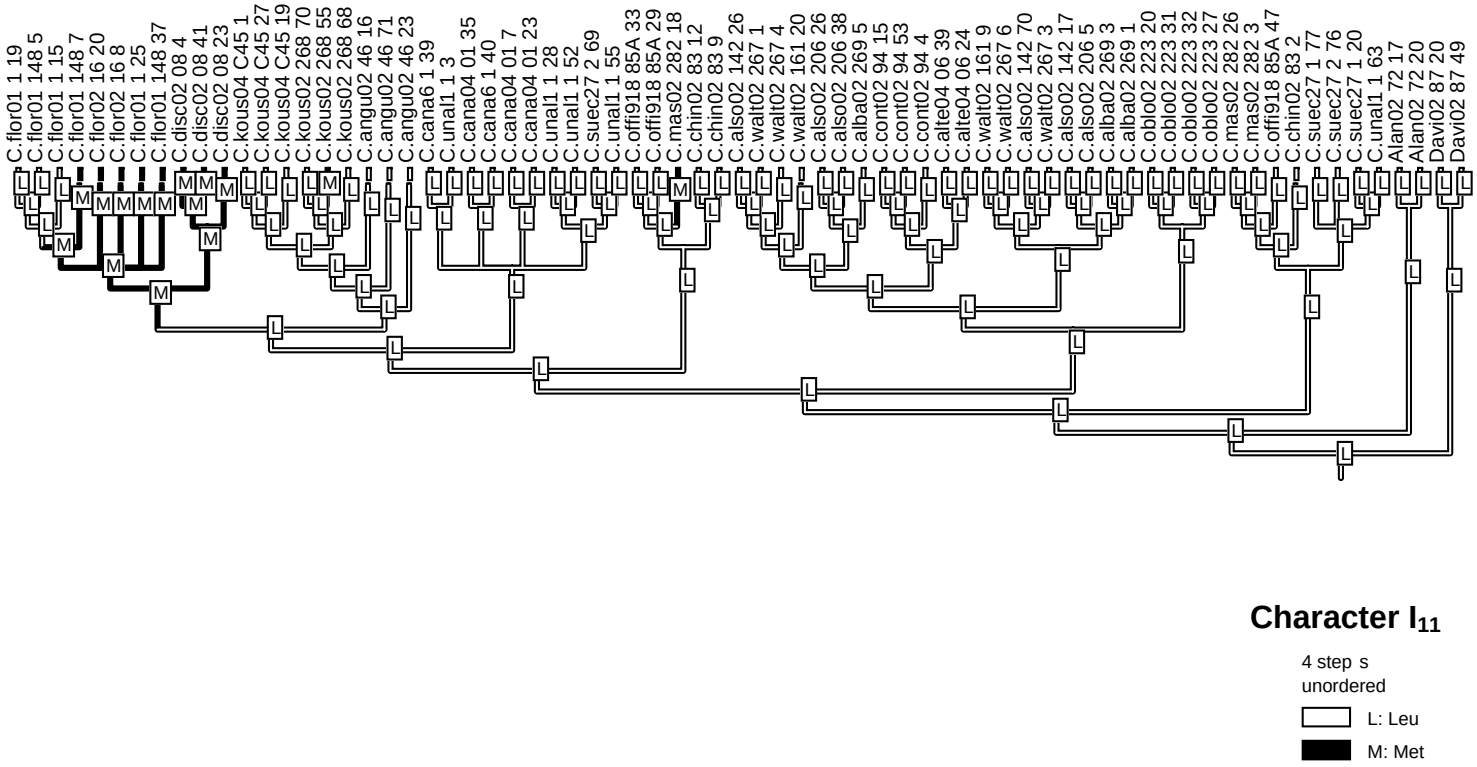
Supplementary materials: character M_{54}



Supplementary materials: character I_4



Supplementary materials: character I_6



Supplementary materials: character I_{11}

Chapter II

Molecular evolution of *APETALA3*-like genes in the dogwood genus *Cornus* (Cornaceae)

Keywords: *APETALA3*-like genes, *Cornus*, gene duplication, MADS-box genes, molecular evolution, phylogeny, regulatory gene.

Abbreviations

AP3: *APETALA3*

CorAP3: *AP3*-like genes in *Cornus*

Abstract

The *APETALA3* (*AP3*) gene encodes MADS domain-containing transcription factors and is responsible for specifying the identity of petals and stamens in *Arabidopsis*. Understanding the evolutionary patterns and processes of MADS-box genes is an important step toward unraveling the molecular basis of floral morphological evolution. However, the evolutionary history, pattern, and processes of MADS-box genes in most flowering plants remain unknown. In this study, I investigate *AP3*-like gene evolution in the dogwood genus *Cornus* (Cornaceae). An *AP3* gene genealogy based on genomic DNA sequences reveals multiple sequence types within many *Cornus* species suggesting that frequent, independent, and recent gene duplications occurred within species. Two divergent copies of the *AP3* gene are found in *Davidia,* a close relative of *Cornus*. The dN/dS ratios based on both MEGA and PAML analyses reveal accelerated evolutionary rates (dN/dS ratios) of the gene in *Cornus* compared to those reported for other MADS-box genes from other flowering plants, suggesting relaxed selective constraints in the *AP3* -like genes of *Cornus*. Despite this relaxed constraint, no positively selected amino acid sites are detected. Selection on *AP3* genes among *Cornus* subgroups is also not significantly different. My results provide little evidence that the rapid evolution in *CorAP3* gene of *Cornus* is driven by the diversifying positive selection.

Introduction

In the last 15 years, genetic studies have made increasingly significant contributions to our growing understanding of the molecular basis of floral organ development and flower evolution. From the landmark ABC model to the extended 'quartet model', five classes of homeotic genes, most of which belong to MADS box gene family were found to function in a combinatorial manner to specify regional identities in the four floral whorls (Pelaz et al. 2000; Theissen and Saedler 2001). The A class genes alone specify sepals; the A, B and E class genes together specify petals; the B, C and E class genes together determine stamens; the C and E class genes determine carpels; while the D and E class genes together specify ovules (Theissen 2001). Evolutionary developmental studies also demonstrate functional divergence of the ABC program across angiosperms (Kramer and Irish 1999, 2000; Theissen et al. 2000), suggesting that floral identity genes and the modification of flower programs may be the key factor in floral morphological evolution. Recent studies in monocots (e.g. rice and maize) show that the ABC model appears to be generally conserved across angiosperms (Ambrose et al. 2000; Whipple et al. 2004), suggesting sequence changes and functional evolution at these gene loci may play critical role in producing floral organ diversity. Therefore, characterization of the pattern of floral homeotic gene evolution through comparative study should provide an important context for understanding morphological change in flowering plants.

The B-class genes, *PISTILLATA* and *APETALA3* are known to form heterodimers that function in regulating the development of petals and stamens in the model plant *Arabidopsis* (Schwarz-Sommer et al. 1992; Tröbner et al. 1992; Goto and Meyerowitz 1994; Riechmann

et al. 1996a; Riechmann et al. 1996b). The B-class genes also have recently been studied in some basal eudicots (e.g. order Ranunculales) and basal angiosperms, which are among the best studied floral organ identity genes from the evolutionary point of view (Irish 2003; Kramer and Jaramillo 2005). Multiple copies of the B-class genes were detected in some of these plants, with gene duplication events at various phylogenetic levels (order, family to intra-specific) suggesting a dynamic pattern of gene evolution (Stellari et al. 2004, Kramer et al. 1998; Kramer and Irish 2000; Kramer et al. 2003). My study of the evolution of *PI*-like genes in *Cornus* also demonstrates frequent gene duplication at various phylogenetic levels (chapter 1). I detect ancient duplication events which predated the initial diversification of *Cornus* and subsequent losses of gene copies in different subgroups during the early radiation of the genus. Each species analyzed is found to contain multiple copies with most of them derived recently within species, suggesting frequent, independent, and recent gene duplications in all species, similar to the finding from studies of basal eudicots, e.g., Ranunculaceae (Kramer et al. 2003). To obtain a more complete picture of B-class gene evolution, I extend our comparative investigations to *AP3*-like genes of *Cornus.*

Recent studies suggest that morphological evolution is largely linked to gene regulation, and that changes in regulatory genes may play a key role in morphological innovation (Doebley 1993; Doebley and Lukens 1998; Purugganan 1998). Major lines of evidence leading to such hypotheses include elevated ratios of replacement to non-replacement substitution rates (measured by ω=dN/dS) in regulatory genes, compared to structural genes, and diversifying or directional selection detected in regulatory gene coding regions (Gaut and Doebley 1997; Purugganan 1998; Barrier et al. 2001; Remington and Purugganan 2002).

However, the relative importance of regulatory genes in contributing to morphological evolution remains unclear. Furthermore, questions remain regarding the major evolutionary forces governing regulatory gene evolution, in particular, whether the accelerated ratio of dN/dS in regulatory genes is due to positive selection or neutral evolution. The floral homeotic genes are an important group of regulatory genes specifying floral organ identities. Investigating the molecular evolution of these genes would shed important light on the molecular basis underlying the evolution and development of floral morphology during the radiation of flowering plants (Purugganan et al. 1995; Purugganan 1998). However, the evolutionary pattern and forces governing gene evolution are largely unknown in most flowering plants. A recent study detected adaptive sequence evolution of the MADS-box gene family in *Arabidopsis*, and this evidence suggested that sites under positive selection in the MADS-box gene family may have played important roles during MADS-box gene diversification to acquire novel functions, as well as during phenotypic evolution of plants (Martinez-Castilla and Alvarez-Buylla 2003). These findings need investigation in diverse plant lineages to advance our understandings of the evolutionary forces governing floral homeotic gene evolution.

In the present study, I report results of comparative genomic sequence analysis of the *AP3*-like genes in the dogwood genus, *Cornus,* an Asterid clade of higher eudicots in the angiosperm tree of life (APGII 2003). Dogwoods (*Cornus*, Cornaceae) radiated into four morphologically distinct subgroups in the early Tertiary (at least 50-60 million years ago based on fossil data) that are diverse in their inflorescence architecture and bract morphology (Eyde 1988; Xiang et al. 2006). Flowers of all dogwood species produce normal flowers

(with sepals, petals, stamens, and pistil), except in one species in which stamens and pistils become sterile on separate plants (a reproductive system of functional dioecy). Two of the four subgroups (e.g., the flowering dogwood, *Cornus florida*; and the bunch berry, *C. canadensis,* and their allies) evolved four petaloid bracts at the base of their inflorescences, while the bracts in other subgroups are not petaloid (leafy or scale-like). The showy bracts have been the key trait giving the flowering dogwoods (*C. florida*, *C. kousa*) enormous horticulture values. The genus has a well-established phylogeny from previous studies of multiple genes (Xiang et al. 1993, 1996, 1998, 2002, 2006; Fan and Xiang 2001, 2003), which provide an important framework for elucidating gene evolution. The subgroups of the genus are old enough to allow observation of sequence divergence of coding regions, but likely not too old to erase substitutions at the synonymous sites due to mutation saturation. Through analyses of genomic DNA sequences with multiple sampling from each subgroup and its constituent species, I provide a detailed genealogy of the *AP3*-like genes in the genus to study the following aspects regarding this gene: 1) examining the rate and pattern of the *AP3* gene evolution via reconstructing an explicit evolutionary history (genealogy) of the gene from genomic DNA data, and 2) determining the evolutionary force acting on the *AP3* gene in *Cornus*.

Materials and Methods

Taxon Sampling

Cornus consists of ~55 species divided among four major lineages. The blue- or white-fruited group (BW) is the most diverse and produces open compound cymes with minute,

early deciduous green bracts. The cornelian cherry group (CC) has 6 species, producing umbellate cymes with four basal, scale-like bracts. The big-bracted group (BB) contains ~8-13 species that possess capitate cymes subtended by four large, petaloid bracts except one species (*C. disciflora*) which lacks the petaloid bracts. The last group is the dwarf dogwoods (DW), which is the only herbaceous lineage in the genus with 3 species that bear minute compound cymes subtended by four petaloid bracts. I include samples of 15 *Cornus* species representing all major lineages and four morphological subgroups of the genus (Table 1, Fan and Xiang 2001; Xiang et al. 2006). Sampling of multiple species from the same subgroup is employed to confirm and detect variation in the copy number of *CorAP3* within subgroups. Two genera, *Alangium* (Alangiaceae) and *Davidia* (Davidiaceae), close relatives of *Cornus* based on phylogenetic analyses of Cornales (Xiang et al. 1998, 2002; Fan and Xiang 2003) are used as outgroups (Table 1). *Alangium* is the sister genus of *Cornus*, and *Davidia* is in a clade sister to the *Alangium-Cornus* clade (Xiang et al. 1998, 2002; Fan and Xiang 2003). *AP3* homologs of other eudicots (*Arabidopsis thaliana* AY142590, *Hydrangea macrophylla* AF230702, and *Antirrhinum majus* X52023), and *AP3*-like genes of lower eudicots and basal angiosperms (*Sanguinaria canadensis* AF130868, *Dicentra eximia* AF052875, *Meliosma dilleniifolia* AY436709, *Calycanthus floridus* AF230700, *Thottea siliquosa* AY436716, *Houttuynia cordata* AB089154, *Lindera erythrocarpa* AY436736, *Chloranthus spicatus* AY397762, *Nymphaea tetragona* AB158351) are obtained from Genbank for a broader investigation of the pattern of *AP3* homolog diversification in flowering plants in comparison to those observed within *Cornus*.

CorPI Cloning and Sequencing

CorAP3 loci are amplified from total genomic DNA extracted from leaves using the Qiagen DNeasy extraction kit (Qiagen, Hilden, Germany) or the modified CTAB miniprep method (Culings et al. 1992; Xiang et al. 1998). A fragment of approximately 600 base pairs is amplified using *Cornus*-specific *AP3*-like gene PCR primers, which are located on the two conserve regions of MADS-box domain (exon 1) and K domain (exon 4) (Figure 1). This region corresponds to the same region examined for the *PI*-like gene (see Chapter 1). These primers (Forward primer: CAP3-1L2 5'GAATGAGCTCACCGTTCTTTGCGA3' and Reverse primer: CAP3-KR 5'CAAGTCCTCATAGCTCAGATCGTTCA3') are designed based on cDNA sequences of *AP3*-like genes from *Cornus florida* (BB group) and *Cornus alba* (BW group) (provided by Drs. Jer-Ming Hu and Michael Frohlich unpublished data). The PCR reaction contains 5μL of 10X Mg^{2+} free buffer (Promega), 6μL of 25mmol/L $MgCL_2$ (Promega), 10μL of 2.5mmol/L of each dNTPs, 1μL of 20μmol/L forward primer, 1μL of 20μmol/L reverse primer, 5μL of BSA (Bovine serum albumin, 10mg/ml), 1.5 units of Taq polymerase (Promega), 5-10μL of 20ng/μL total DNA extract, and calibrated to a final volume of 50μL using sterile deionized water. A hot-start step, 6 min of 96 ^{o}C incubation, is processed to denature the genomic DNA templates before adding Taq polymerase (Promega). PCR is then performed under 94 ^{o}C for 1 min followed by 36 cycles of 94 ^{o}C (30 s), 63 ^{o}C (1 min), 72 ^{o}C (2 min) and a final step of 7 min extension at 72 ^{o}C. PCR products are gel purified using the Qiaquick gel extraction kit (Qiagene). Direct sequencing of PCR products produce unreadable polymorphic sequences suggesting that multiple copies of the gene co-amplified in most samples. Direct sequencing of a single *AP3*-like gene copy is only possible in *C. officinalis* and *Alangium* spp.

To separate allelic and/or paralogous versions of the gene, a cloning procedure is employed. PCR products are isolated and cleaned using the QIAquick gel extraction kit (Qiagen, USA) to increase ligation efficiency before applying the TOPO TA cloning kit with competent cells following the manufacturer's protocol (Invitrogen). Positive transformants are detected by insert size through PCR screening using CAP3-1L2 and CAP3-KR primers. For all cloned samples, 40 PCR products of positive transformants from each sample (Table 1) are cut with restriction endonucleases Taq I (Promega) and RsaI (New England BioLabs) to detect different sequence types for sequencing. The restriction endonucleases are chosen based on sequence data of 6 clones from *C. florida* (Voucher 01-148), which cleave differently among sequence types, thus generating different cutting patterns for different copies of *AP3*-like genes. A total of one to three positive clones representing each banding pattern in the RFLP analysis are inoculated in a nutrient medium to multiply the cells. Plasmid DNA with *AP3*-like gene inserts are extracted and purified using the Minipreps DNA purification system (Promega) for subsequent sequencing analyses (PerkinElmer). Forward and reverse sequencing are performed to control error in Taq polymerase (PerkinElmer) replication of the target sequence using CAP3-1L2 forward and CAP3-KR reverse primers. In all, 125 clones are sequenced and all sequences are deposited in Genbank.

Phylogenetic analyses

Clone sequences from each sample are first analyzed using the Bellerophon program (Huber et al 2004) to detect chimeric sequences. Three sequences are found to be chimeric

possibly formed through PCR recombination. These chimeric sequences are excluded from subsequent analyses.

An initial phylogenetic analysis is performed by including all *AP3*-like genes of *Cornus* and other flowering plants (see Materials and Methods, taxon sampling). Exon regions alone are used in this analysis because introns could not be reliably aligned between sequences from Cornaceae s. l. (*Cornus*, *Alangium*, and *Davidia* included in this study) and other families. A total of 81 sequences 218 bp in length are included in this analysis. Phylogenetic analyses of this matrix are conducted using Maximum Parsimony (MP), Neighbor-Joining (NJ), and Bayesian Metropolis-Hastings coupled Markov chain Monte-Carlo (MCMC) methods to reconstruct gene genealogy. MP analysis is carried out using PAUP* 4.0b10 (Swofford 2002) using the heuristic search algorithm with branch swapping performed using the tree-bisection-reconnection procedure. The PAUP MULPARS option is in effect. In all, five parsimonious trees are found and these are summarized in the strict consensus tree presented in Figure 2. Node support is assessed by 100,000 replicate bootstrap searches using the fast heuristic search option. For both NJ and Bayesian analyses, Modeltest 3.06 (Posada and Crandall, 1998) is performed to select a best-fitting model of sequence evolution. Neighbor-Joining analysis (Saitou and Nei 1987) under the Jukes-Cantor model is performed using PAUP* 4.0b10 (Swofford 2002) with 10,000 bootstraps. For Bayesian analysis, MCMC phylogenetic analyses carry out using MrBayes 3.1.2 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) with flat priors and three heated chains in addition to a single cold chain. MCMC analyses are conducted in two parallel runs simultaneously initiated with a random tree, and run for a total of 1,000,000 generations (sampling trees

every 100 generations). Both searches reached stationarity after approximately 100,000 generations detected using Tracer v1.1 (Rambaut and Drummond. 2004). The first 1,000 trees are discarded as burn-in and are not included in generating the consensus phylogeny. Summary statistics and consensus phylograms with nodal posterior probability support are estimated from the combination of trees after burn-in. In both neighbor-joining and Bayesian analyses, the trees are rooted using lower eudicots and basal angiosperms (Fig 2).

Phylogenetic analyses of *CorAP3* sequences including both intron and exon sequences are also conducted. Sequences are aligned using Clustal X (Thompson et al. 1997) and adjusted by eye. Based on the Modeltest results, the best fitting model for this dataset is TIM + I + Γ. Both NJ and Bayesian analyses are performed as described above.

Estimates of selection

A total of 38 distinct coding sequences are included in selection analyses, covering the complete I domain and partial MADS- and K- domains of the gene (Fig 4). First, I test for differential selection pressure along branches. Two hypotheses are examined: (1) do *CorAP3* genes differ in evolutionary rate when compared to their close relatives in *Alangium* and *Davidia*? (2) do the *AP3*-like genes in the four different morphological subgroups of *Cornus* evolve at different rates? For the pairwise comparison analyses, I estimate nonsynonymous (dN) and synonymous (dS) nucleotide substitutions according to the method of Nei and Gojobori (1986) using the Jukes-Cantor model, as implemented in MEGA 3.1 (Kumar et al. 2004). Pairwise dN/dS values are calculated for all *CorAP3* sequences, and each of the four major subgroups of *Cornus* (Fig 5). Significance of differences in dN/dS values among

lineages are examined by using two-sample bootstrap resampling test with 10,000 replicates. The dN/dS values are then plotted against the dS values to detect the pattern of dN/dS ratio variation in relation to dS (Fig 6).

To further examine selection pressure on *AP3* in cornalean lineages (e.g., between *Cornus* and its outgroup genera, or among morphological subgroups), I perform analyses applying various codon-based likelihood models in the Codeml program implemented in PAML 3.15 (Yang 1997). The Bayesian tree topology is applied as a test topology in PAML analyses (Fig 7). Branch lengths applied in subsequent analyses are estimated assuming a single dN/dS ratio for all sites (Goldman and Yang 1994). To estimate dN and dS values for each branch in the tree, we apply a model that allows all 59 branches of the phylogeny to have different parameters. To test for diversifying selection between *Cornus* and its outgroups *Alangium* and *Davidia*, I define model A allowing different dN/dS between *Cornus* and its outgroups. A likelihood ratio test is used to determine significance of dN/dS values between lineages by comparing the likelhood values obtained under each model to the likelihood value obtained under a null model (model N) of a single dN/dS value. Similarly, model B allows different dN/dS values among the four *Cornus* morphological subgroups. The likelihood ratio test (LRT) compares twice the difference in log-likelihood values to a χ^2 distribution, with degrees of freedom equal to the difference in the number of free parameters between the two models (Yang 1998). We also estimated the dN/dS ratios among the four morphological groups of *Cornus* using a codon based likelihood approach in PAML for comparison with results obtained from the pairwise comparison method.

We performed further analyses to test for adaptive molecular evolution by examining if any amino acid sites are under positive selection. The coding sequences of 35 *CorAP3* genes of *Cornus* are applied for this analysis. Six codon-substitution models implemented in Codeml in PAML 3.15 (Yang 1997, Yang et al. 2000, Table 2) are applied in the analysis. I compared the models M0 (one-ratio model, ω values are the same at all amino acid sites), M1 (neutral model, all amino acid sites are either completely constrained ω = 0 or neutral ω = 1), and M7 [beta model, ω values are limited to the interval (0, 1) and eight categories of sites with independent ω distributed according to a ß distribution], all three representing the null hypotheses of no positively selected sites allowed with models M2 (selection model, similar to M1 with an additional category of amino acid sites under positive selection ω > 1), M3 (discrete model, three categories of unconstrained ω of amino acid sites were estimated from the data), and M8 (beta & ω model, adds an extra site category to M7 with a free ω ratio estimated from the data allowing ω >1) representing the alternative hypotheses that allow for heterogeneous and positive selection on amino acid sites (Yang et al. 2000). By comparing these nested models using the LRT (i.e., M0 vs. M3, M1 vs. M2, and M7 vs. M8), we could determine whether models allowing for positive selection fit the data better indicating positive selection on the gene, and which amino acid sites are under positive selection (Table 2 and 3, Yang et al. 2000). All analyses for detecting positively selected amino acid sites were performed using Bayes empirical analysis (Yang et al. 2005) implemented in Codeml in PAML 3.15 (Yang 1997).

Rate of molecular evolution and divergence time estimation

I employ the penalized likelihood (PL) method in r8s 1.71 (Sanderson 2002) to estimate divergence times and absolute substitution rate over the entire *CorAP3* gene, including both coding and noncoding regions at all nodes of the gene genealogy. Branch lengths are generated using maximum likelihood (ML) in PAUP* 4.0b10 (Swofford 2002) using the tree topology found by both NJ and Bayesian analyses. Fossil information of *Cornus* is used to place calibration points on the genealogy. Because the age of *Cornus* lineage is hypothesized to be less than 86 mya, I fix the *Cornus* clade at maximum age of 80 mya. In r8s, a cross-validation is performed to determine the optimal value of the rate smoothing parameter (Sanderson 2002). If smoothing is set to be large, then the model is reasonably clocklike; if it is small, then rate variation is allowed (Sanderson 2002). Normally, the smoothing parameter is between 1 and 1000. The estimation of absolute divergence time and rate estimate had been done by using penalized likelihood method with TN algorithm suggested by the method (Sanderson 2002).

Results

Chimeric clones and Taq polymerase error

PCR recombination and Taq polymerase error are potential sources of sequence artifacts in the PCR-based cloning approach applied here to identify unique sequence types. Three clones are identified as chimeric and excluded from later analyses. Taq error in PCR occurs at a rate (mutation frequency/bp/duplication) of $8 * 10^{-6}$ (Cline et al. 1996). For a 600 bp sequence in this study, there should be less than 1 (actually value 0.29) error expected. Thus,

sequence difference among clones examined can be inferred as having arisen from alleles, orthologs, or paralogs.

Structural characteristics and sequence variation

The novel sequences obtained in this study are confirmed to be plant *AP3*-like genes using BlastN search (Altschul et al. 1997). The sequenced region of *Cornus AP3*-like genes ranges from 512 to 595 bp among the 15 sampled species. This region spans the exons of partial MADS-box, complete I- and partial K- domains and the introns between them (Fig 1). The exon regions are conservative in size among species. The coding regions covered 216 bp in this study including 51 bp out of the 171 bp of the MADS domain, the complete 93 bp of the I-domain, and 72 bp out of the 198 bp of the K-domain. However, the intron regions vary greatly in length among species of *Cornus*, ranging from 84 to 137 bp for intron 1, 85 to 125 bp for intron 2, and 95 to145 bp for intron 3.

All the exon sequences are easily aligned among *Cornus* and outgroup taxa without introducing indels. The exons from *Cornus* species alone contain 22.22% (48/216) parsimony informative sites and 37.50% (81/216) variable sites. When considering Cornaceae s. l. (including *Alangium*, *Davidia,* and *Hydrangea*) or other eudicots and basal angiosperms as well as Cornaceae s. l. (see materials and methods), these numbers increase to 57.41% (124/216) and 78.24% (169/216) for variable sites and 42.13% (91/216) and 58.33% (126/216) for parsimony informative sites, respectively. The matrices including both exon and intron sequences are constructed for *CorAP3* in *Cornus* and its outgroup genera *Davidia* and *Alangium*. This matrix contains 727 bp aligned sequences, of which 55.71%

(405/727) are parsimony-informative, and 64.51% (469/727) are variable. When only *Cornus* species are included, there are 37% (269/727) parsimony-informativesites and 46% (338/727) variable.

Genealogy and gene copies

All phylogenetic analyses (MP/NJ/Bayesian) indicate that the sequence variants of *AP3*-like genes obtained in *Cornus* can be sorted into clades corresponding to the major morphological subgroups, except that *C. oblonga* is unexpectedly excluded from the BW clade (Fig 2). Moreover, two divergent *AP3*-like gene copies are detected in *Davida*, a monotypic genus (Fig 2). The expected relationships of Cornaceae s. l. (((*Cornus*, *Alangium*), *Davidia*), *Hydrangea*) based on previously published molecular data (Xiang et al. 1998, 2002; Fan and Xiang 2003) is not recovered by analysis of *AP3* coding sequences. This is may be due to the limited number of informative sites available, or due to the possible loss of the *Cornus*-orthologous copy of the gene in *Alangium*, or simply due to insufficient sampling from *Alangium*. *Cornus* is a moderately supported clade in this analysis, although resolution among subgroups is lacking. All subgroups of *Cornus* are well supported, except for BB which has weak support. The multiple sequence types detected within each species form groups indicating that all gene duplications likely occurred after the origins of sampled species. The single 4n species *C. unalaschkensis* is more closely related to *C. canadensis*, however the node support and posterior probability for this relationship are weak.

Phylogenetic analyses including both coding and non-coding regions of *CorAP3* resulted in strongly supported and concordant relationships in both Bayesian and NJ analyses

(Fig 3). These results agree with coding sequence analyses (Fig 2) in supporting the monophyly of *CorAP3* lineages and the monophyly of *AP3* from each major morphological subgroup. The BB lineage, which is weakly supported in the analysis of coding sequence alone, is strongly supported when noncoding regions are included. However, relationships between the major subgroups of *Cornus* are still poorly resolved and the *AP3*-like gene from *C. oblonga* is placed outside of the BW clade. In the DW group, all sequences from the 4n species, *C. unalaschkensis,* are embedded within the 2n *C. canadensis* clade, indicating that the latter might be a progenitor of the 4n species.

Multiple sequence types of *CorAP3* are detected within most of the examined species of *Cornus*. At least 2 different sequence types are observed in all species, except that in *C. mas* and *C. officinalis*, only one sequence type is detected (Fig 3). Sequence types show differences in nucleotide substitution, but also length variation in introns. Sequence variants are more divergent in a single sample than expected from Taq error alone. However, it is not possible for us to determine whether these represent multiple gene copies or allelic differences. The number of alleles cannot exceed two in each plant. Therefore, plant samples with more than two sequence types must have at least one additional copy of the gene. Given that sequence types from the same species are grouped together in trees (Fig 2 and Fig 3), my results suggest that gene duplications occurred within species independently.

Rate of substitution, and selection of *CorAP3*

Pairwise synonymous substitution rates (dS) of *CorAP3* estimated in MEGA range from 0.02 to 0.39 with a mean of 0.18. These values are approximately 3 times higher than

pairwise nonsynonymous substitution rates (dN), which range from 0 to 0.13 with an average of 0.06. The mean dN/dS ratio of the gene is 0.33 ± 0.16. Among four morphological subgroups, the mean dN/dS values vary from 0.42 ± 0.32 (in the BW including *C. oblonga*, 0.36 ± 0.37 excluding *C. oblonga*) to 0.27 ± 0.21 (in the BB), with values for the CC and DW groups being intermediate (0.34 ± 0.31 and 0.32 ± 0.17, respectively) (Fig 5). For outgroups, dS values are about 3 times greater than dN values (dS: 0.37 - 0.77, average 0.63; and dN: 0.11 - 0.22, average 0.17). The mean dN/dS ratio for *Davidia* and *Alangium* is lower than for *Cornus* (0.28 ± 0.03 vs. 0.33± 0.16). However, a 2-sample bootstrap resampling test indicates no statistical significance for dN/dS ratio differences between outgroup species and *Cornus* (P = 0.47) and among major subgroups (P-value, 0.06 - 0.88). The dN/dS ratios between all pairs of *CorAP3* sequences are mostly below 1 (98.3%, dN/dS = 0 - 1) except for a few pairs which are slightly above 1 (1.7%, dN/dS = 1 - 1.5) (Fig 5, *Cornus CorAP3*). Plots of dN/dS versus dS indicate that the ratio of dN/dS is negatively related to dS for all categories analyzed (Fig 6), suggesting that the *CorAP3* gene is under strong functional constraint, and the rate of nonsynonymous changes is constrained as sequences diverged.

The codon-based likelihood method (Yang 1997) indicates that 53 sites (24.5%) are synonymous substitution and 163 sites (75.5%) are nonsynonymous substitution among total 216 bp coding region. When I allow estimates to change freely along all branches, dS values vary from 0 to 0.28, dN values range from 0 to 0.07, and dN/dS ratios vary from 0.0001 to 256.42 excluding branches with dS equal to zero. Approximately one third of branches have extremely large dN/dS values due to the zero value of dS (see supplementary materials). There are 27% (16 out of 59) branches having a dN/dS value over 1, and the maximum

dN/dS value is 257.32 and the minimum is 0.000l (see supplementary materials). Among these branches with dN/dS ratio over 1, two have a dS > 0, suggesting the large dN/dS ratios of these branches observed are not artifacts from a zero value of dS. For all branches with zero dS, four have ≥ 2 mutations at nonsynonymous sites, a greater number than expected from equal rate of dN, dS and an indication of dN/dS ratio truly greater than 1 on those branches. These branches include terminal branches leading to the herbaceous species *C. canadensis* 04-01-6 and *C. oblonga* 02-223-15, an enigmatic BW species, an internal branch leading to *C. alsophila* 02-143-17 and *C. alsophila* 02-143-34 (supplementary materials), and one leading to one of two copies found in *Davida* 02-87-2. For the first three branches metioned above, the nonsynonymous change is slightly greater than 2 while synonymous change equals zero, suggesting that diversifying or positive selection occurred on these branches. Intriguingly, the branch leading to *Davidia* 02-87-2 has 12.2 nonsynonymous changes while synonymous change is 0, indicating strong positive selection on this gene copy, possibly for divergence in function. The high level of sequence divergence observed between these two copies of *AP3*-like gene in *Davidia* may reflect an ancient duplication of *AP3*-like genes in *Davidia*. The gene genealogy suggests that this duplication may have occurred in the common ancestor of *Cornus*, *Alangium* and *Davidia* groups, with subsequent loss of a copy in ancestors of *Cornus* and *Alangium* (Fig 3).

For the PAML analysis, when different dN/dS ratios of *CorAP3* in *Cornus* and *AP3*-like genes in outgroup species are allowed, the estimate of dN/dS for *CorAP3* is 0.26 which is 1.5 times larger than the dN/dS ratio of 0.17 in outgroups (Fig 7). However, the likelihood value for this model is -1192.60, which is not significantly different than the null model (Ln = -

1193.18; LRT, P = 0.28). This suggests that the difference in dN/dS values between *CorAP3* and *AP3*-like genes of outgroups is not statistically significant, similar to the finding from analyses based on data from pairwise comparisons of MEGA. Comparisons of dN/dS values among four morphological subgroups of *Cornus* showed there are substantial differences in the selection force among subgroups, with the highest dN/dS value in the BW group (0.34, excluding *C. oblonga*), the least ratio in the BB group (0.14), and intermediate ratios in the DW and CC groups (0.31 and 0.24, respectively) (Fig 7). These estimates also agree with the pairwise comparison results (Fig 5). The likelihood ratio test indicates there is not significant difference (P = 0.56) between the model that allows different dN/dS values among subgroups (Ln = -1191.21) and the null model (Ln = -1193.18) that assumes the same dN/dS ratios among *Cornus* lineages. This suggests no difference of selection pressure exists among morphological subgroups.

Under various codon-based substitution models for testing positive selection at amino acid sites (Yang et al., 2000), the estimates of average dN/dS values of *CorAP3* rang from 0.22 to 0.28 among all models (Table 2), indicating that the purifying selection dominates the evolution of the gene. The one-ratio model (M0) is easily rejected when compared with all other models allowing dN/dS ratio to vary among sites. None of the models that allows for the presence of positively selected sites, that is, M2 (selection), M3 (discrete), and M8 (beta & ω), detected positively selected amino acid sites on the *CorAP3* gene (Table 2). Only model M3 has the significantly higher likelihood value than models M0, and none of M2 and M8 models is a better fitting my data than their counterparts M1 and M7 which do not allow for positive selection (Table 3).

Absolute rates and divergence time estimation

Absolute substitution rates of *CorAP3* reveal substantial differences among four major lineages of *Cornus*. When the value of the smoothing parameter is 3 giving the best cross-validation score for *CorAP3* data, this suggest that the best model for the data deviates from clocklike rate constancy and high rate variation of the gene exists among *Cornus* lineages (Sanderson 2002). The substitution rate in the DW ($2.2*10^{-3} \sim 4.0*10^{-3}$ substitutions/site/mya) is obviously elevated when compared to other species of *Cornus* ($2.7*10^{-4} \sim 1.8*10^{-3}$) (Figure 8, and supplementary material).

Discussion

AP3*-like gene evolution in *Cornus

In core eudicots, most species investigated to date have only one copy of *AP3*-like genes, except for *Brassica oleraceae* (Brassicaceae) and *Rumex acetosa* (Polygonaceae) which have two copies (Kramer et al. 1998). Phylogenetic reconstruction indicates that those *AP3* paralogs were formed by an independent gene duplication event within the species or its genus (Kramer et al. 1998). Here, I find that the *CorAP3* lineage in *Cornus* does not experience ancient gene duplications. However, multiple sequence types of *CorAP3* genes are detected in most of the species of *Cornus*, suggesting gene duplication occurred independently within these species. In addition, I find two divergent copies of *AP3*-like genes in *Davidia,* the close relative of *Cornus*, which produces petaloid bracts and lacks petals and sepels. These findings agree with recent studies of B-class gene evolution, which suggested a

dynamic and stochastic pattern of gene evolution (Stellari et al. 2004, Kramer et al. 1998; Kramer and Irish 2000; Kramer et al. 2003).

The *CorAP3* gene genealogy resolves phylogenetic relationships of species within each *Cornus* subgroup, and also strongly supports the monophyly of *Cornus*, however the relationships among the morphological subgroups are poorly resolved (Fig 2 and 3). This lack of resolution among major lineages of *Cornus* is likely due to the low number of informative sites available. The other explanation for this observation could be that *CorAP3* genes may have experienced an extremely low rate of change during the early diversification of *Cornus*. Alternatively, rates of molecular evolution may be heterogeneous among lineages and over time, so it is possible that *CorAP3* was carried through a rapid radiation of *Cornus* species early in the history of the group - a finding is supported by hypotheses of relationship based on other markers (Xiang et al. 2006), and in the *CorPI* tree (Fig 2 and 3, Charpter 1).

The nucleotide substitution rates of *CorAP3* genes in DW lineage are about two to ten times larger than other subgroups (Fig 8 and supplementary materials). The accelerated nucleotide substitution rate in the DW clade is also observed in other molecular markers (Xiang et al. 2006) and in *CorPI*-A (Chapter 1). This phenomenon may be related to the generation time of DW dogwoods, as DW species are herbaceous with short generation times.

Functional constraints and heterogeneoity of nucleotide substitution rates

It is generally believed that regulatory genes evolve faster than structural protein genes (e.g., Purugganan and Wessler 1994; Purugganan 1998; Ting et al. 1998; Barrier et al., 2001).

The dN/dS values of *CorAP3* are 0.33 ± 0.16 (0.26 in PAML) and 0.28 ± 0.03 (0.17) in outgroup genera *Alangium* and *Davidia* (Fig 5 and 7). These values are greater than the mean dN/dS value (0.14) detected for other plant nuclear loci. Furthermore, they are also higher than 0.11-0.19 in MADS-box genes and 0.12 ± 0.03 for *AP3*-like genes found in other species (Purugganan et al. 1995), which suggests that selection on these regulatory genes was relaxed in *Cornus* and its relatives. Even though these values are significantly lower than found for their functional counterpart *CorPI* (average dN/dS ratio is 0.49 ± 0.38; Chapter 3).

Similarly, as in *CorPI*, the dN/dS values of *CorAP3* genes are generally less than 1 (Table 2 and Fig 5) and the dN/dS ratio is negatively related to dS (Fig 6) indicating that purifying selection on this gene. My data suggest functional constraints on the gene restricting substitutions at nonsynonymous sites. However, at least a few branches with large dN/dS values have a dS>0 or dN greater than expected based on number of dN and dS sites (Supplementary materials). Extremely large dN/dS values detected on certain branches may be an indication of positive selection acting on those branches. However, I can not exclude the possibility of bias that while nonsynonymous sites changed, no synonymous sites had yet accumulated at these terminal branches. These branches with extremely large values of dN/dS indicate that some sites or certain region of the gene are under strong positive selection in these lineages. In particularly, the outgroup species, *Davidia,* there are 12.2 nonsynonymous changes leading to one paralogous copy 02-87-2, while zero synonymous changes were found on that same branch, which provides strong evidence of diversifying positive selection acting on this gene copy after gene duplication. Strong positive selection

detected in one of the *AP3* copies in *Davidia* suggests that this paralogous gene copy is likely diverging for new function after gene duplication.

It is expected that most proteins are under purifying selection with a dN/dS ratio close to 0, because most amino acids in the protein are under strong functional constraints (Li et al. 1985). The elevated dN/dS ratio of *CorAP3* genes in *Cornus* is probably due to relaxation of selection. This relaxed constraint resembles complete neutrality at ~ 16-18% of amino acid sites (Table 2). This is consistent with the notion that most amino acid substitutions in this gene are due to genetic drift, and the average ω does not differ across lineages (see results). My findings agree with recent studies of regulatory genes that no positive selection is found for ortholog genes, *Ipmyb1* in *Ipomoea* (Chang et al. 2005), and *GA1* in the Hawaiian silversword alliance (Remington and Purugganan 2002), suggesting positive selection may correlate with gene duplication and functional divergence of paralogous genes.

Acknowledgement

We thank Drs. Jer-Ming Hu and Michael Frohlic who provided primers and mRNA sequence data of *C. florida* and *C.alba* for this study. We are also grateful to the NSF supported DEEPTIME program (funded to D. E. Soltis DEB-0090283) for the travel support to workshops on divergence time dating and application of fossil data. We thank M. D. Purugganan, B. M. Wiegmann, and N. S. Allen for the insightful comments on the early version of the manuscript. We thank D. Thomas and K. M. Ding for help with collecting data.

This study is supported by National Science Foundation grants to Q-Y (J.) X (NSF-DEB 0129069 and DEB 0444125).

References:

ALTSCHUL, S. F., T. L. MADDEN, A. A. SCHAFFER, J. ZHANG, Z. ZHANG, W. MILLER, AND D. J. LIPMAN. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25: 3389-3402.

AMBROSE, B. A., D. R. LERNER, P. CICERI, C. M. PADILLA, M. F. YANOFSKY, AND R. J. SCHMIDT. 2000. Molecular and genetic analyses of the *silky1* gene reveal conservation in floral organ specification between eudicots and monocots. Molecular Cell 5: 569-579.

APG II (ANGIOSPERM PHYLOGENY GROUP). 2003. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants. Botanical Jorunal of the Linnean Society 141: 399-436.

BARRIER, M., R. H. ROBICHAUX, AND M. D. PURUGGANAN. 2001. Accelerated regulatory gene evolution in an adaptive radiation. Proceedings of the National Academy of Sciences of the United States of America 98: 10208-10213.

CHANG, S. M., Y. LU, AND M. D. RAUSHER. 2005. Neutral evolution of the nonbinding region of the anthocyanin regulatory gene *Ipmyb1* in *Ipomoea*. Genetics 170: 1967-1978.

CLINE, J., J. C. BRAMAN, AND H. H. HOGREFE. 1996. PCR fidelity of *pfu* DNA polymerase and other thermostable DNA polymerases. Nucleic Acids Research 24: 3546-3551.

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

DOEBLEY, J. 1993. Genetics, development and plant evolution. Current Opinion in Genetics and Development 3: 865-872.

DOEBLEY, J. AND L. LUKENS. 1998. Transcriptional regulators and the evolution of plant form. Plant Cell 10: 1075-1082.

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

FAN, C., M. D. PURUGGANAN, D. T. THOMAS, B. M. WIEGMANN, AND J. Q. XIANG. 2004. Heterogeneous evolution of the *Myc*-like Anthocyanin regulatory gene and its phylogenetic utility in *Cornus* L. (Cornaceae). Molecular Phylogenetics and Evolution 33: 580-594.

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

-----. 2003. Phylogenetic analyses of Cornales based on 26S rDNA and combined 26S rDNA-*matK-rbcL* sequence data. American Journal of Botany 90: 1357-1372.

GAUT, B. S. AND J. F. DOEBLEY. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. Proceedings of the National Academy of Sciences of the United States of America 94: 6809-6814.

GOLDMAN, N. AND Z. H. YANG. 1994. A codon-based model of nucleotide substitution for protein-coding DNA-sequences. Molecular Biology and Evolution 11: 725-736.

GOTO, K. AND E. M. MEYEROWITZ. 1994. Function and regulation of the Arabidopsis floral homeotic gene *PISTILLATA*. Genes & Development 8: 1548-1560.

HUBER, T., G. FAULKNER, AND P. HUGENHOLTZ. 2004. Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. Bioinformatics 20: 2317-2319.

HUELSENBECK, J. P. AND F. RONQUIST. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17: 754-755.

IRISH, V. F. 2003. The evolution of floral homeotic gene function. Bioessays 25: 637-646.

KRAMER, E. M., V. S. DI STILIO, AND P. M. SCHLUTER. 2003. Complex patterns of gene duplication in the *APETALA3* and *PISTILLATA* lineages of the Ranunculaceae. International Journal of Plant Sciences 164: 1-11.

KRAMER, E. M., R. L. DORIT, AND V. F. IRISH. 1998. Molecular evolution of genes controlling petal and stamen development: duplication and divergence within the *APETALA3* and *PISTILLATA* MADS-box gene lineages. Genetics 149: 765-783.

KRAMER, E. M. AND V. F. IRISH. 1999. Evolution of genetic mechanisms controlling petal development. Nature 399: 144-148.

-----. 2000. Evolution of the petal and stamen developmental programs: Evidence from comparative studies of the lower eudicots and basal angiosperms. International Journal of Plant Sciences 161: S29-S40.

KRAMER, E. M. AND M. A. JARAMILLO. 2005. Genetic basis for innovations in floral organ identity. Journal of Experimental Zoology (Molecular and Developmental Evolution) 304B:526-535.

KUMAR, S., K. TAMURA, AND M. NEI. 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Briefing in Bioinformatics 5: 150-163.

LI, W. H., C. I. WU, AND C. C. LUO. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Molecular Biology and Evolution 2: 150-174.

MARTINEZ-CASTILLA, L. P. AND E. R. ALVAREZ-BUYLLA. 2003. Adaptive evolution in the Arabidopsis MADS-box gene family inferred from its complete resolved phylogeny. Proceedings of the National Academy of Sciences of the United States of America 100: 13407-13412.

NEI, M. AND T. GOJOBORI. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular Biology and Evolution 3: 418-426.

PELAZ, S., G. S. DITTA, E. BAUMANN, E. WISMAN, AND M. F. YANOFSKY. 2000. B and C floral organ identity functions require *SEPALLATA* MADS-box genes. Nature 405: 200-203.

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

PURUGGANAN, M. D. 1998. The molecular evolution of development. Bioessays 20: 700-711.

PURUGGANAN, M. D., S. D. ROUNSLEY, R. J. SCHMIDT, AND M. F. YANOFSKY. 1995. Molecular evolution of flower development: diversification of the plant MADS-box regulatory gene family. Genetics 140: 345-356.

PURUGGANAN, M. D. AND S. R. WESSLER. 1994. Molecualr evolution of the plant *R* regulatory gene family. Genetics 138: 849-854.

RAMBAUT, A. AND A. J. DRUMMOND 2004. Tracer. University of Oxford, Oxford.

REMINGTON, D. L. AND M. D. PURUGGANAN. 2002. *GAI* homologues in the Hawaiian silversword alliance (Asteraceae-Madiinae): molecular evolution of growth regulators in a rapidly diversifying plant lineage. Molecular Biology and Evolution 19: 1563-1574.

RIECHMANN, J. L., B. A. KRIZEK, AND E. M. MEYEROWITZ. 1996a. Dimerization specificity of Arabidopsis MADS domain homeotic proteins *APETALA1*, *APETALA3*, *PISTILLATA*, and *AGAMOUS*. Proceedings of the National Academy of Sciences of the United States of America 93: 4793-4798.

RIECHMANN, J. L., M. WANG, AND E. M. MEYEROWITZ. 1996b. DNA-binding properties of Arabidopsis MADS domain homeotic proteins *APETALA1*, *APETALA3*, *PISTILLATA* and *AGAMOUS*. Nucleic Acids Research 24: 3134-3141.

RONQUIST, F. AND J. P. HUELSENBECK. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572-1574.

SAITOU, N. AND M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4: 406-425.

SANDERSON, M. J. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Molecular Biology and Evolution 19: 101-109.

SCHWARZ-SOMMER, Z., I. HUE, P. HUIJSER, P. J. FLOR, R. HANSEN, F. TETENS, W. E. LONNIG, H. SAEDLER, AND H. SOMMER. 1992. Characterization of the *Antirrhinum* floral homeotic MADS-box gene *deficiens*: evidence for DNA binding and autoregulation of its persistent expression throughout flower development. EMBO Journal 11: 251-263.

STELLARI, G. M., M. A. JARAMILLO, AND E. M. KRAMER. 2004. Evolution of the *APETALA3* and *PISTILLATA* lineages of MADS-box-containing genes in the basal angiosperms. Molecular Biology and Evolution 21: 506-519.

SWOFFORD, D. L. 2002. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4.10b. Sinauer Associates, Sunderland, Mass.

THEISSEN, G. 2001. Development of floral organ identity: stories from the MADS house. Current Opinion in Plant Biology 4: 75-85.

THEISSEN, G., A. BECKER, A. DI ROSA, A. KANNO, J. T. KIM, T. MUNSTER, K. U. WINTER, AND H. SAEDLER. 2000. A short history of MADS-box genes in plants. Plant Molecular Biology 42: 115-149.

THEISSEN, G. AND H. SAEDLER. 2001. Plant biology. Floral quartets. Nature 409: 469-471.

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 25: 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.

TROBNER, W., L. RAMIREZ, P. MOTTE, I. HUE, P. HUIJSER, W. E. LONNIG, H. SAEDLER, H. SOMMER, AND Z. SCHWARZ-SOMMER. 1992. *GLOBOSA*: a

homeotic gene which interacts with *DEFICIENS* in the control of *Antirrhinum* floral organogenesis. EMBO Journal 11: 4693-4704.

WHIPPLE, C. J., P. CICERI, C. M. PADILLA, B. A. AMBROSE, S. L. BANDONG, AND R. J. SCHMIDT. 2004. Conservation of B-class floral homeotic gene function between maize and *Arabidopsis*. Development 131: 6083-6091.

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

XIANG, J. Q., M. L. 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, J. Q., 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, J. Q., 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.

XIANG, Q. Y. J., D. T. THOMAS, W. H. ZHANG, S. R. MANCHESTER, AND Z. MURRELL. 2006. Species level phylogeny of the genus *Cornus* (Cornaceae) based on

molecular and morphological evidence - implications for taxonomy and Tertiary intercontinental migration. Taxon 55: 9-30.

YANG, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Computer applications in the biosciences 13: 555-556.

-----. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Molecular Biology and Evolution 15: 568-573.

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.

YANG, Z., W. S. WONG, AND R. NIELSEN. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. Molecular Biology and Evolution 22: 1107-1118.

Table 1. Source of plant materials used in this study and information of clones analyzed. Specimens used in this study including taxa represented, reference identifiers, vouchers. Voucher specimens were collected by Xiang and deposited in NCSC.

Subgroups†	Species	Voucher and collection locality	Abbreviation in analyses	No. of clones analyzed	No. of clones sequenced
BW	*C. alba* L.	02-269, Beijing China	Calba02-269	40	11
	C. alsophila W.W.Sm.	02-143, Yunnan, China	Calso02-143	40	5
	C. controversa Hemsl.	02-94, Sichuan, China	Ccont02-94	40	6
	C. oblonga Wall.	02-223, Yunnan, China	Coblo02-223	40	8
	C. walteri Wangerin	02-267, Beijing, China	Cwalt02-267	40	8
CC	*C. chinensis* Wangerin	02-83, Sichuan, China	Cchin02-83	40	5
	C. mas L.	02-282, JC Raulston Arboretum	Cmas02-282	40	7
	C. officinalis Siebold & Zucc.	918-85A, Arnold Arboretum	Coffi918-85A	40	4

Table 1. continued.

Subgroups†	Species	Voucher and collection locality	Abbreviation in analyses	No. of clones analyzed	No. of clones sequenced
BB	*C. angustata* (Chun)T.R.Dudley	02-46, Guangxi, China	Cangu02-46	40	9
	C. disciflora Sesse & Moc. ex DC.	02-08, Heredia, Costa Rico	Cdisc02-08	40	8
	C. florida L.	01-148, North Carolina, USA	Cflor01-148	40	7
	C. kousa Buerger ex Miq.	04-C45, Sichuan, China	Ckous04-C45	40	6
DW	*C. canadensis* L.	04-01, West Virginia, USA	Ccana04-01	40	10
	C. suecica L.	Pop27-2, Alaska, USA	Csuec27-2	40	10
	C. unalaschkensis Ledeb.	Pop4-1, British Columbia, Canada	Cunal4-1	40	11
Outgroups	*Alangium sp.*	02-72, Guangxi, China	Alan02-72	-	1*
	Davidia involucrate Baill.	02-87, Sichuan, China	Davi02-87	40	9

†BW: blue- or white- fruited group; CC: cornelian cherry; BB: big-bracted group; DW: dwarf dogwoods. *: *Alangium* has only one sequence type of *AP3* homologs.

Table 2. Parameters estimated for the *APETALA3*-like genes of *Cornus* based on various codon-based substitution models implemented in PAML 3.15 to test selection.

Model	p[a]	lnL[b]	dN/dS	Estimates of Parameters[c]	Positively Selected Sites[d]
M0: one-ratio	1	-1193.18	0.224	ω=0.224	None
M1: neutral (K=2)	1	-1183.61	0.280	P_0=0.844	None
M2: selection (K=3)	3	-1183.61	0.280	P_0=0.844, P_1=0.156, (P_2=0), ω_2=15.53	None
M3: discrete (K=3)	5	-1183.13	0.253	P_0=0.078, P_1=0.745, (P_2=0.176), ω_0=0, ω_1=0.152, ω_2=0.792	None
M7: beta	2	-1184.35	0.242	P=0.818, q=2.526	None
M8: beta&ω	4	-1184.50	0.270	P_0=0.870, (P_1=0.130), P=2.885, q=14.859, ω=1.000	None

a: Number of parameters in the model.

b: Log-likelihood scores; also see Yang et al. (2000) for the definitions of parameters.

c: P_i denotes the proportion of site falling in site class ω_i (see Yang et al. 2000 for explanation of parameters).

d: Sites potentially under positive selection identified under model M2, M3 and M8 with a posterior probability *: >95% or **: >99%.

Table 3. Comparisons of models (from Table 2) based on Likelihood Ratio Tests (LRTs) for testing heterogeneous selection at amino acid sites in *CorAP3* genes.

LRT	Degrees of freedom	$\chi 2$ Critical Value (5%)	$\chi 2$ Critical Value (1%)	$2*\Delta l$
M0 vs. M3	4	9.49	13.28	20.10**
M1 vs. M2	2	5.99	9.21	0.96
M7 vs. M8	2	5.99	9.21	1.70

(*: P>95%; **: P>99%)

Supplementary material. Parameters estimated for selection pressure based on codon-based likelihood method implemented in PAML 3.15. Model allows variable ω among all branches.

branch	t	N	S	dN/dS	dN	dS	N*dN	S*dS
39..40	0.346	162.9	53.1	0.0883	0.0326	0.3688	5.3	19.6
40..41	0.035	162.9	53.1	0.0001	0	0.0481	0	2.6
41..42	0.042	162.9	53.1	0.6557	0.0125	0.0191	2	1
42..1	0.014	162.9	53.1	202.846	0.0062	0	1	0
42..2	0.014	162.9	53.1	0.0001	0	0.0193	0	1
41..43	0.028	162.9	53.1	0.3308	0.0062	0.0188	1	1
43..3	0	162.9	53.1	0.3007	0	0	0	0
43..4	0.014	162.9	53.1	0.0001	0	0.0187	0	1
41..44	0.014	162.9	53.1	0.0001	0	0.019	0	1
44..5	0.043	162.9	53.1	0.0001	0	0.058	0	3.1
44..45	0	162.9	53.1	0.3987	0	0	0	0
45..6	0.014	162.9	53.1	0.0001	0	0.0189	0	1
45..7	0.014	162.9	53.1	205.8183	0.0062	0	1	0
40..46	0.096	162.9	53.1	0.1379	0.0126	0.0916	2.1	4.9
46..47	0	162.9	53.1	0.5488	0	0	0	0
47..8	0	162.9	53.1	0.1752	0	0	0	0
47..9	0.028	162.9	53.1	0.0001	0	0.0381	0	2
47..10	0.014	162.9	53.1	204.5755	0.0063	0	1	0
46..48	0.014	162.9	53.1	206.0716	0.0063	0	1	0
48..49	0.014	162.9	53.1	0.0001	0	0.0189	0	1
49..11	0.014	162.9	53.1	205.2848	0.0064	0	1	0
49..12	0.014	162.9	53.1	0.0001	0	0.0188	0	1
48..50	0	162.9	53.1	0.5478	0	0	0	0
50..13	0	162.9	53.1	0.3032	0	0	0	0
50..14	0.014	162.9	53.1	205.639	0.0063	0	1	0
40..51	0.13	162.9	53.1	0.6605	0.0384	0.0581	6.3	3.1
51..52	0.028	162.9	53.1	0.0001	0	0.0386	0	2
52..15	0.014	162.9	53.1	204.0003	0.0062	0	1	0
52..16	0	162.9	53.1	0.1102	0	0	0	0
51..53	0.014	162.9	53.1	0.0001	0	0.0188	0	1
53..54	0	162.9	53.1	0.3966	0	0	0	0
54..17	0	162.9	53.1	0.153	0	0	0	0
54..18	0.014	162.9	53.1	0.0001	0	0.0188	0	1
53..19	0.014	162.9	53.1	205.5888	0.0062	0	1	0
51..55	0	162.9	53.1	0.3997	0	0	0	0
55..56	0	162.9	53.1	0.3997	0	0	0	0
56..20	0.014	162.9	53.1	0.0001	0	0.0189	0	1

56..21	0	162.9	53.1	0.1393	0	0	0	0
56..22	0.028	162.9	53.1	0.3297	0.0062	0.0189	1	1
55..57	0.029	162.9	53.1	255.2441	0.0127	0	2.1	0
57..23	0.015	162.9	53.1	0.0001	0	0.0197	0	1
57..24	0.029	162.9	53.1	0.3181	0.0063	0.0198	1	1.1
40..58	0.083	162.9	53.1	0.3579	0.0191	0.0533	3.1	2.8
58..25	0.029	162.9	53.1	257.3202	0.0128	0	2.1	0
58..26	0.014	162.9	53.1	0.0001	0	0.0193	0	1
58..27	0	162.9	53.1	0.0001	0	0	0	0
40..59	0.355	162.9	53.1	0.2317	0.0653	0.2817	10.6	15
59..28	0.015	162.9	53.1	204.937	0.0065	0	1.1	0
59..29	0.03	162.9	53.1	256.4193	0.013	0.0001	2.1	0
59..30	0.015	162.9	53.1	205.6863	0.0065	0	1.1	0
59..31	0.014	162.9	53.1	203.9549	0.0062	0	1	0
59..32	0.014	162.9	53.1	0.0001	0	0.0192	0	1
59..60	0.029	162.9	53.1	0.0001	0	0.0391	0	2.1
60..33	0.028	162.9	53.1	0.3212	0.0062	0.0193	1	1
60..34	0.014	162.9	53.1	203.6028	0.0062	0	1	0
60..35	0	162.9	53.1	0.0958	0	0	0	0
39..36	0.498	162.9	53.1	0.1034	0.053	0.513	8.6	27.2
39..37	0.171	162.9	53.1	92.3629	0.0751	0.0008	12.2	0
39..38	1.349	162.9	53.1	0.1106	0.1512	1.3663	24.6	72.5

t: branch length; N: number of nonsynonymous substitution sites; S: number of synonymous substitution sites; dN: nonsynonymous substitution rate; dS: synonymous substitution rate; dN/dS: ratio of nonsynonymous substitution rate versus synonymous substitution rate; N*dN: number of nonsynonymous substitution changes; S*dS: number of synonymous substitution changes.

Supplementary material. Estimations of divergence times and nucleotide substitution rates for each node and each branch of the phylogeny (Fig 8) based on *CorAP3* coding and noncoding regions using r8s 1.71.

Nodes	Age	ESR	Nodes	Age	ESR	Nodes	Age	ESR
1	662.03	-	C.suec27 2 10	0	3.89E-03	C.also02 143 17	0	8.78E-04
2	214.91	1.15E-03	C.suec27 2 17	0	4.04E-03	C.also02 143 34	0	5.63E-04
Fixed	80	1.34E-03	C.cana04 01 11	0	2.79E-03	48	15.12	6.75E-04
5	75.67	1.74E-03	C.unal4 1 18	0	2.76E-03	49	12.62	5.17E-04
6	17.76	1.21E-03	C.cana04 01 6	0	2.27E-03	C.alba02 269 2	0	1.26E-04
7	16.41	1.03E-03	C.cana04 01 39	0	2.44E-03	C.alba02 269 3	0	7.14E-04
8	3.46	8.63E-04	C.unal4 1 7	0	2.15E-03	C.alba02 269 4	0	5.64E-04
C.flor01 148 15	0	5.98E-04	31	29.67	1.09E-03	53	5.02	1.09E-03
C.flor01 148 7	0	1.01E-03	32	3.93	6.64E-04	C.cont02 94 1	0	1.35E-03
11	6.9	1.01E-03	C.chin02 83 54	0	4.55E-04	C.cont02 94 8	0	5.33E-04
C.disc02 08 20	0	8.73E-04	C.chin02 83 74	0	8.61E-04	56	13.45	1.04E-03
C.disc02 08 25	0	1.12E-03	C.chin02 83 55	0	4.55E-04	57	7.11	1.19E-03
14	14.15	1.15E-03	36	9.26	1.19E-03	C.walt02 267 17	0	9.25E-04
15	7.77	1.09E-03	37	2.82	1.37E-03	C.walt02 267 9	0	1.48E-03
C.kous04 C45 6	0	9.57E-04	C.mas02 282 12	0	1.40E-03	C.walt02 267 38	0	2.72E-04
C.kous04 C45 8	0	1.18E-03	C.mas02 282 14	0	1.40E-03	61	5.42	1.57E-03
C.angu02 46 15	0	1.16E-03	C.offi918 85A 17	0	9.84E-04	C.oblo02 223 15	0	1.30E-03
19	4.38	2.57E-03	41	73.05	1.47E-03	63	2.1	1.78E-03
20	4	2.58E-03	42	41.34	1.38E-03	C.oblo02 223 4	0	1.82E-03
21	3.67	2.69E-03	43	37.82	1.12E-03	C.oblo02 223 5	0	1.82E-03
22	3.42	3.12E-03	44	33.07	8.57E-04	Davi02 87 2	0	1.28E-03
23	1.2	3.84E-03	45	5.74	7.58E-04	Davi02 87 23	0	7.36E-14

Unit for divergence time estimation is million years before present (MYBP), and for estimation of substitution rates (ESR) is substitutions/site/million year. Hypothesized age of the earliest divergence of *Cornus* is fixed at 80 MYBP. Program can not estimate ESR for node 1, the basal branch. Node numbers correspond to internodes in the phylogenetic tree (see Fig 9); sample names indicate terminal branches.

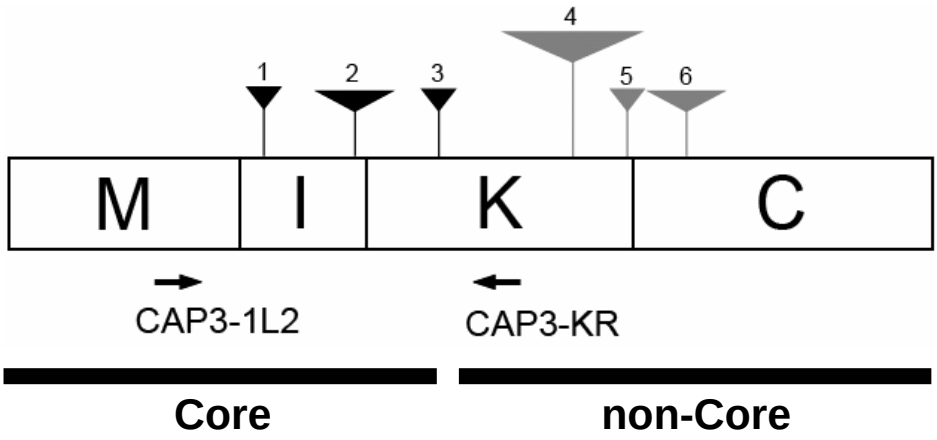


Figure 1. Schematic diagram of MIKC structure of *APETALA3* homologs based on complete sequence of *Antirrhinum major* (X62810). The region between primers CAP3-1L2 and CAP3-KR is analyzed for this study. Reversed triangles indicate position and relative sizes among six introns. Core and non-core regions are indicated based on functional studies.

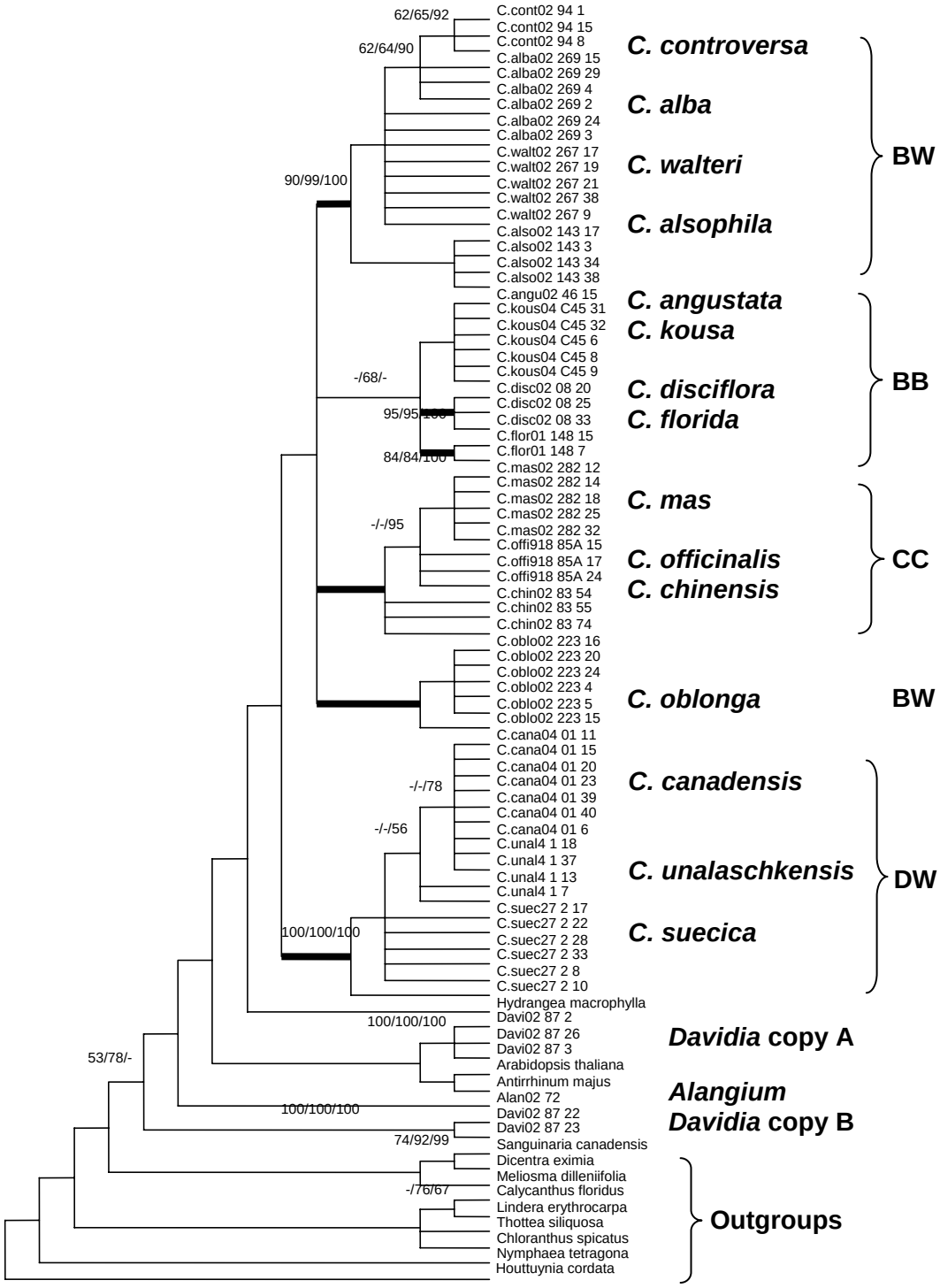

62/65/92
C.cont02 94 1
C.cont02 94 15
62/64/90
C.cont02 94 8
C. controversa
C.alba02 269 15
C.alba02 269 29
C.alba02 269 4
C. alba
C.alba02 269 2
C.alba02 269 24
C.alba02 269 3
C.walt02 267 17
C. walteri
C.walt02 267 19
90/99/100
C.walt02 267 21
C.walt02 267 38
C.walt02 267 9
C. alsophila
C.also02 143 17
C.also02 143 3
C.also02 143 34
C.also02 143 38
BW
C.angu02 46 15
C. angustata
C.kous04 C45 31
C. kousa
C.kous04 C45 32
C.kous04 C45 6
C.kous04 C45 8
-/68/-
C.kous04 C45 9
C.disc02 08 20
C. disciflora
95/95/100
C.disc02 08 25
C.disc02 08 33
C. florida
C.flor01 148 15
84/84/100
C.flor01 148 7
C.mas02 282 12
BB
C.mas02 282 14
C. mas
C.mas02 282 18
C.mas02 282 25
-/-/95
C.mas02 282 32
C.offi918 85A 15
C.offi918 85A 17
C. officinalis
C.offi918 85A 24
C.chin02 83 54
C. chinensis
C.chin02 83 55
C.chin02 83 74
CC
C.oblo02 223 16
C.oblo02 223 20
C.oblo02 223 24
C.oblo02 223 4
C. oblonga
BW
C.oblo02 223 5
C.oblo02 223 15
C.cana04 01 11
C.cana04 01 15
C.cana04 01 20
-/-/78
C.cana04 01 23
C. canadensis
C.cana04 01 39
C.cana04 01 40
-/-/56
C.cana04 01 6
C.unal4 1 18
C.unal4 1 37
C. unalaschkensis
C.unal4 1 13
C.unal4 1 7
DW
C.suec27 2 17
C.suec27 2 22
100/100/100
C.suec27 2 28
C. suecica
C.suec27 2 33
C.suec27 2 8
C.suec27 2 10
Hydrangea macrophylla
100/100/100
Davi02 87 2
Davi02 87 26
Davidia copy A
53/78/-
Davi02 87 3
Arabidopsis thaliana
Antirrhinum majus
Alangium
Alan02 72
100/100/100
Davi02 87 22
Davidia copy B
Davi02 87 23
74/92/99
Sanguinaria canadensis
Dicentra eximia
Meliosma dilleniifolia
-/76/67
Calycanthus floridus
Lindera erythrocarpa
Thottea siliquosa
Outgroups
Chloranthus spicatus
Nymphaea tetragona
Houttuynia cordata

Figure 2. Strict consensus tree of 5 most parsimonious trees from parsimony analysis based on coding sequences of *AP3*-like genes rooted by lower eudicots and basal angiosperms. Numbers indicated near branches are parsimony bootstraps/NJ bootstraps/Bayesian posterior probability, respectively. Hyphens indicate nodes supported by ≤ 50% of bootstrap values and ≤ 60% in Bayesian posterior probability. Bold branches show strongly supported morphological clades. Abbreviations of species names are indicated in Table 1. DW = dwarf dogwoods, CC = cornelian cherry, BW = blue- or white-fruited dogwoods, and BB = big-bracted dogwoods.

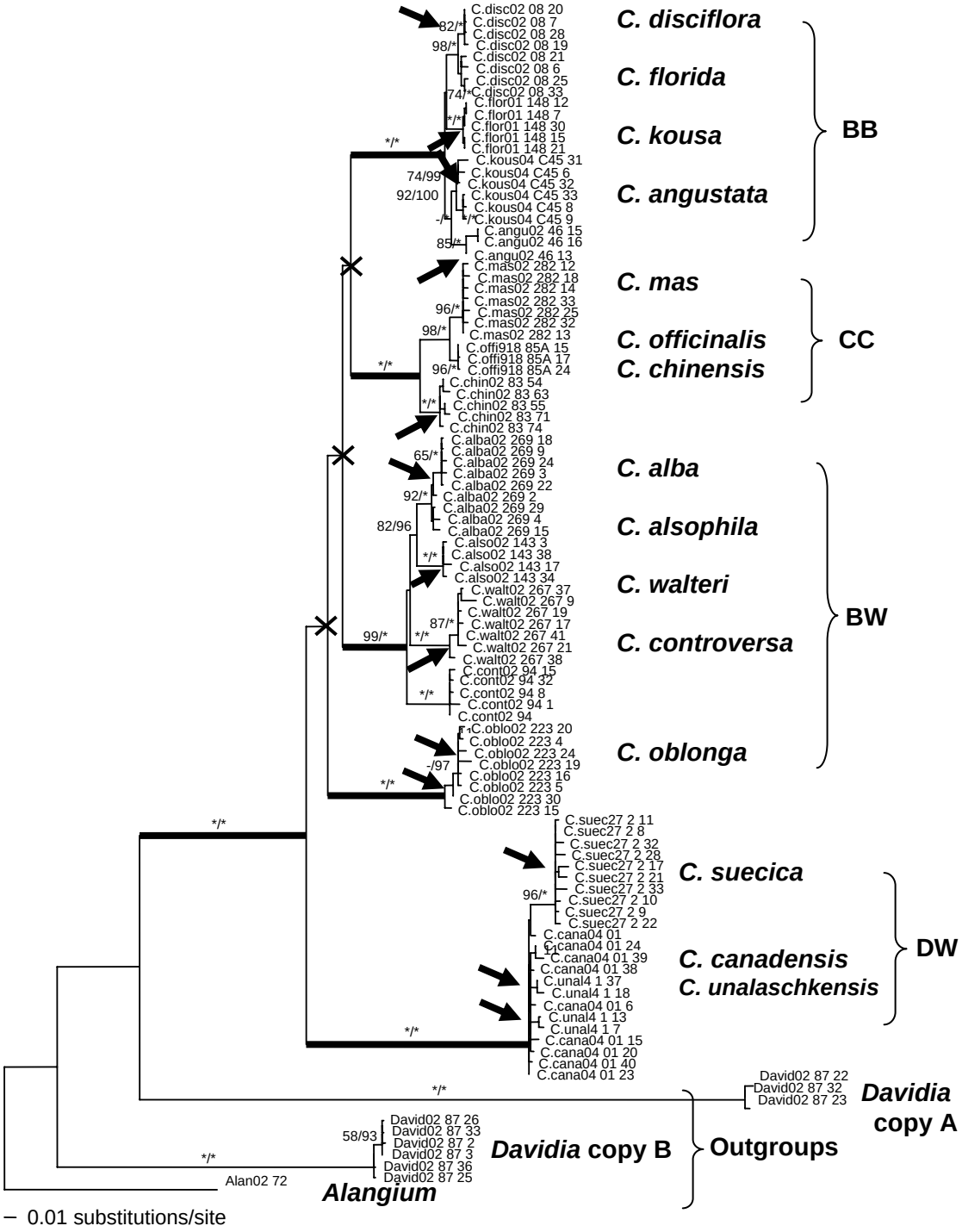

C.disc02 08 20
C.disc02 08 7
C.disc02 08 28
C.disc02 08 19
C.disc02 08 21
C.disc02 08 6
C.disc02 08 25
C.disc02 08 33
C.flor01 148 12
C.flor01 148 7
C.flor01 148 30
C.flor01 148 15
C.flor01 148 21
C.kous04 C45 31
C.kous04 C45 6
C.kous04 C45 32
C.kous04 C45 33
C.kous04 C45 8
C.kous04 C45 9
C.angu02 46 15
C.angu02 46 16
C.angu02 46 13
C. disciflora
C. florida
C. kousa
C. angustata
BB
C.mas02 282 12
C.mas02 282 18
C.mas02 282 14
C.mas02 282 33
C.mas02 282 25
C.mas02 282 32
C.mas02 282 13
C.offi918 85A 15
C.offi918 85A 17
C.offi918 85A 24
C.chin02 83 54
C.chin02 83 63
C.chin02 83 55
C.chin02 83 71
C.chin02 83 74
C. mas
C. officinalis
C. chinensis
CC
C.alba02 269 18
C.alba02 269 9
C.alba02 269 24
C.alba02 269 3
C.alba02 269 22
C.alba02 269 2
C.alba02 269 29
C.alba02 269 4
C.alba02 269 15
C.also02 143 3
C.also02 143 38
C.also02 143 17
C.also02 143 34
C.walt02 267 37
C.walt02 267 9
C.walt02 267 19
C.walt02 267 17
C.walt02 267 41
C.walt02 267 21
C.walt02 267 38
C.cont02 94 15
C.cont02 94 32
C.cont02 94 8
C.cont02 94 1
C.cont02 94
C.oblo02 223 20
C.oblo02 223 4
C.oblo02 223 24
C.oblo02 223 19
C.oblo02 223 16
C.oblo02 223 5
C.oblo02 223 30
C.oblo02 223 15
C. alba
C. alsophila
C. walteri
C. controversa
C. oblonga
BW
C.suec27 2 11
C.suec27 2 8
C.suec27 2 32
C.suec27 2 28
C.suec27 2 17
C.suec27 2 21
C.suec27 2 33
C.suec27 2 10
C.suec27 2 9
C.suec27 2 22
C.cana04 01
C.cana04 01 24
C.cana04 01 39
C.cana04 01 38
C.unal4 1 37
C.unal4 1 18
C.cana04 01 6
C.unal4 1 13
C.unal4 1 7
C.cana04 01 15
C.cana04 01 20
C.cana04 01 40
C.cana04 01 23
C. suecica
C. canadensis
C. unalaschkensis
DW
David02 87 22
David02 87 32
David02 87 23
Davidia copy A
David02 87 26
David02 87 33
David02 87 2
David02 87 3
David02 87 36
David02 87 25
Davidia copy B
Alan02 72
Alangium
Outgroups
0.01 substitutions/site

Figure 3. Phylogram of one *CorAP3* gene tree based on Bayesian analyses including both intron and exon regions. *AP3*-like genes from *Alangium* and *Davidia* are used as outgroups rooting the tree. Numbers near the branches are NJ bootstraps/Bayesian posterior probability, respectively. Stars indicate nodes supported by 100% bootstraps or by 100% Bayesian posterior probabilities. Hyphens indicate nodes supported by ≤ 50% bootstrap values and by ≤ 60% Bayesian posterior probabilities. Bold branches show strongly supported morphological clades. Abbreviations of species names are indicated in Table 1. DW = dwarf dogwoods, CC = cornelian cherry, BW = blue- or white-fruited dogwoods, and BB = big-bracted dogwoods.

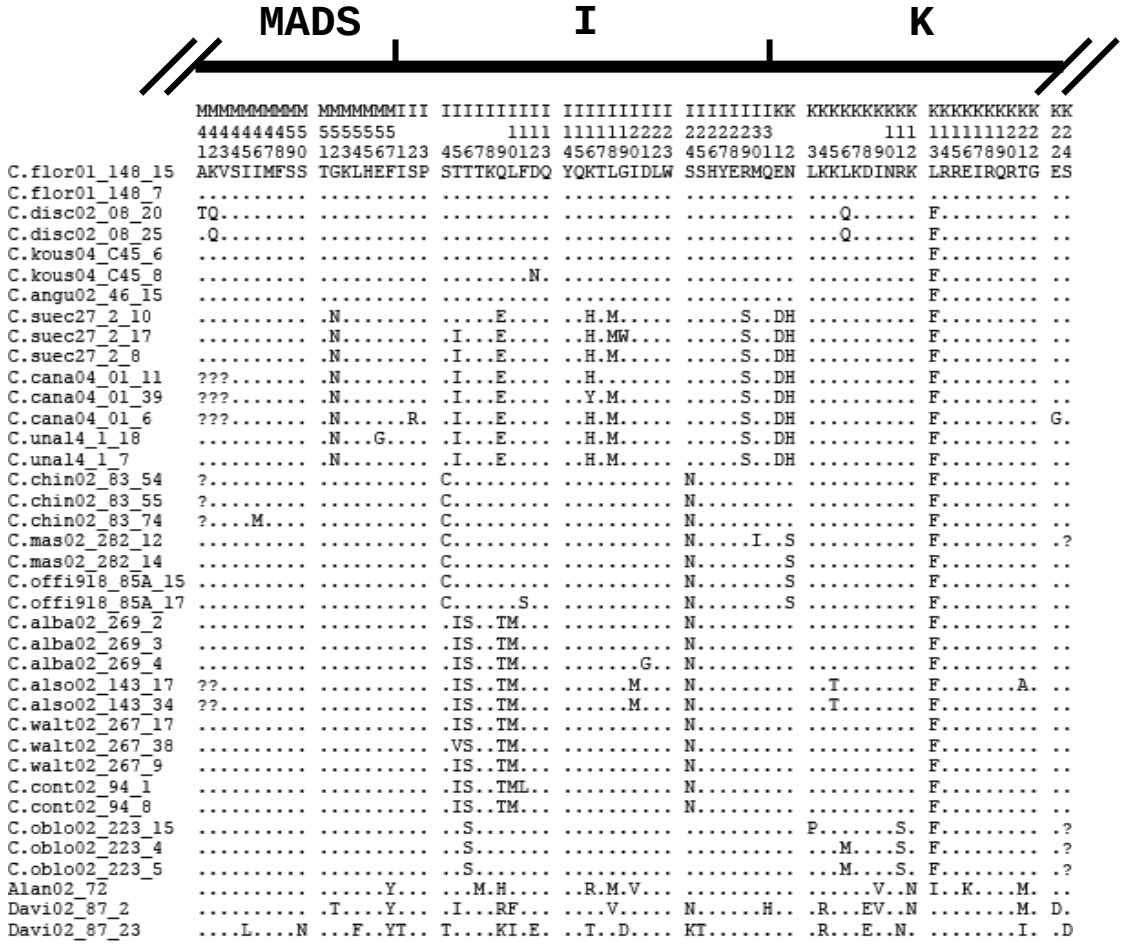
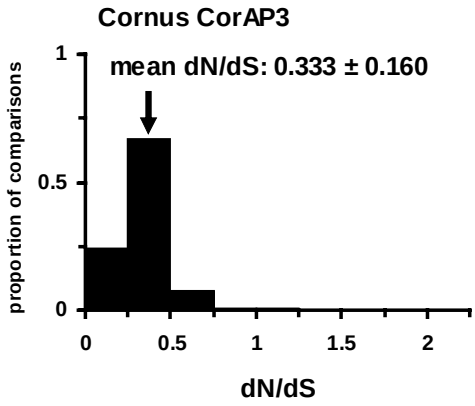
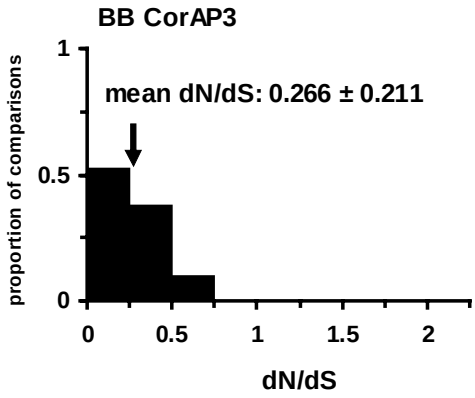
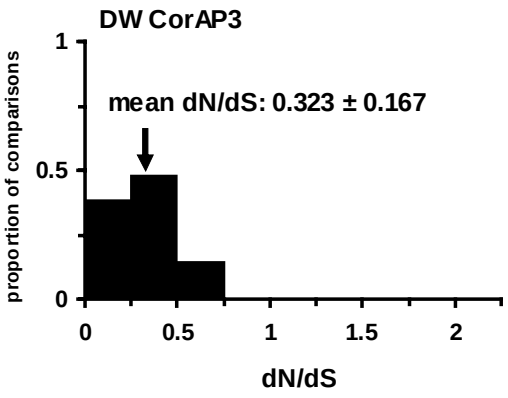
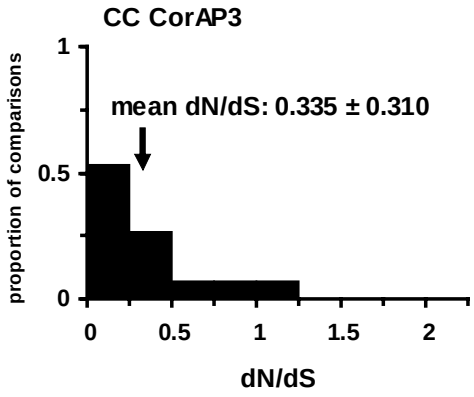


Figure 4. Alignment of coding sequences of 38 *AP3*-like genes from *Cornus* and its outgroup genera *Alangium* and *Davidia* covering whole I-domain and partial MADS- and K-domains, which is applied for MEGA and PAML analyses.


Cornus CorAP3
mean dN/dS: 0.333 ± 0.160
proportion of comparisons
1
0.5
0
0
0.5
1
1.5
2
dN/dS


BB CorAP3
mean dN/dS: 0.266 ± 0.211
proportion of comparisons
1
0.5
0
0
0.5
1
1.5
2
dN/dS


DW CorAP3
mean dN/dS: 0.323 ± 0.167
proportion of comparisons
1
0.5
0
0
0.5
1
1.5
2
dN/dS


CC CorAP3
mean dN/dS: 0.335 ± 0.310
proportion of comparisons
1
0.5
0
0
0.5
1
1.5
2
dN/dS

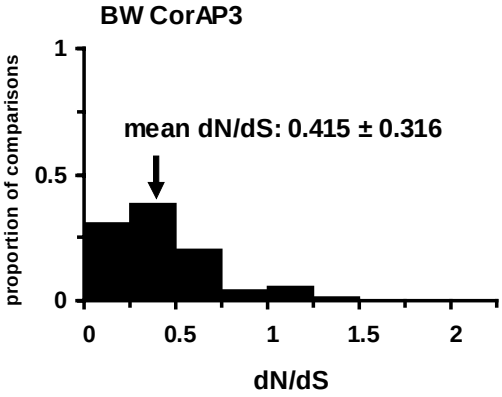
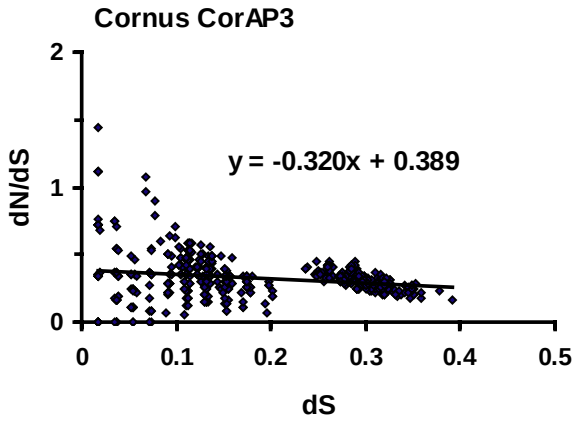
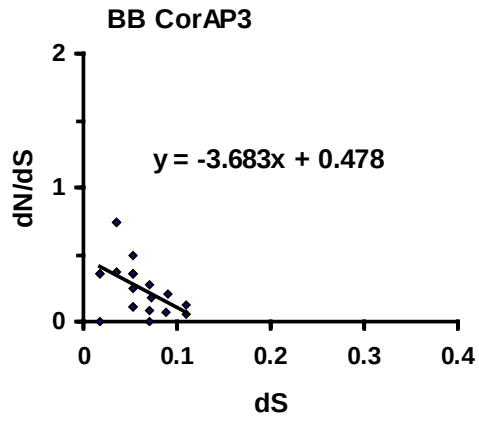
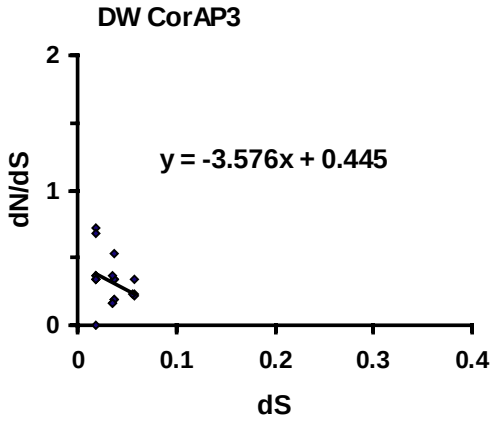
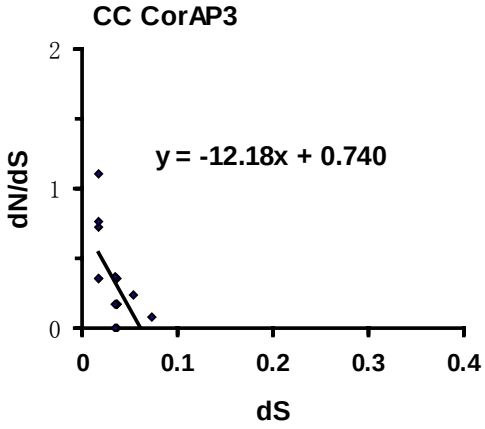

BW CorAP3
mean dN/dS: 0.415 ± 0.316
proportion of comparisons
1
0.5
0
0
0.5
1
1.5
2
dN/dS

Figure 5. Distribution of dN/dS ratios based on pairwise comparisons (modified Nei-Gojobori, Jukes-Cantor model) for all *CorAP3* genes, and for each of four morphological subgroups, with mean values and standard deviations indicated. Pairwise comparisons that had dS = 0 are excluded from analyses. Individual pairwise dN and dS values are estimated from MEGA 3.1.


Cornus CorAP3
y = -0.320x + 0.389
dN/dS
dS
0
1
2
0
0.1
0.2
0.3
0.4
0.5


BB CorAP3
y = -3.683x + 0.478
dN/dS
dS
0
1
2
0
0.1
0.2
0.3
0.4


DW CorAP3
y = -3.576x + 0.445
dN/dS
dS
0
1
2
0
0.1
0.2
0.3
0.4


CC CorAP3
y = -12.18x + 0.740
dN/dS
dS
0
1
2
0
0.1
0.2
0.3
0.4

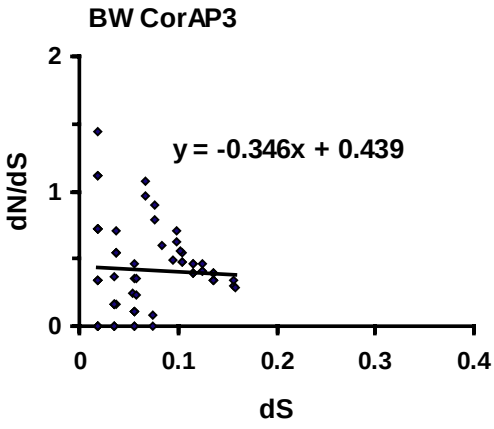

BW CorAP3
y = -0.346x + 0.439
dN/dS
dS
0
1
2
0
0.1
0.2
0.3
0.4

Figure 6. Plots of dN/dS ratios versus dS based on estimates of pairwise comparisons (modified Nei-Gojobori, Jukes-Cantor model) using MEGA 3.1 for all *CorAP3* genes, and for each of four morphological subgroups. Regression lines and equations indicate trend of relationships between dN/dS and dS. Pairwise comparisons that had dS = 0 are excluded from analyses.

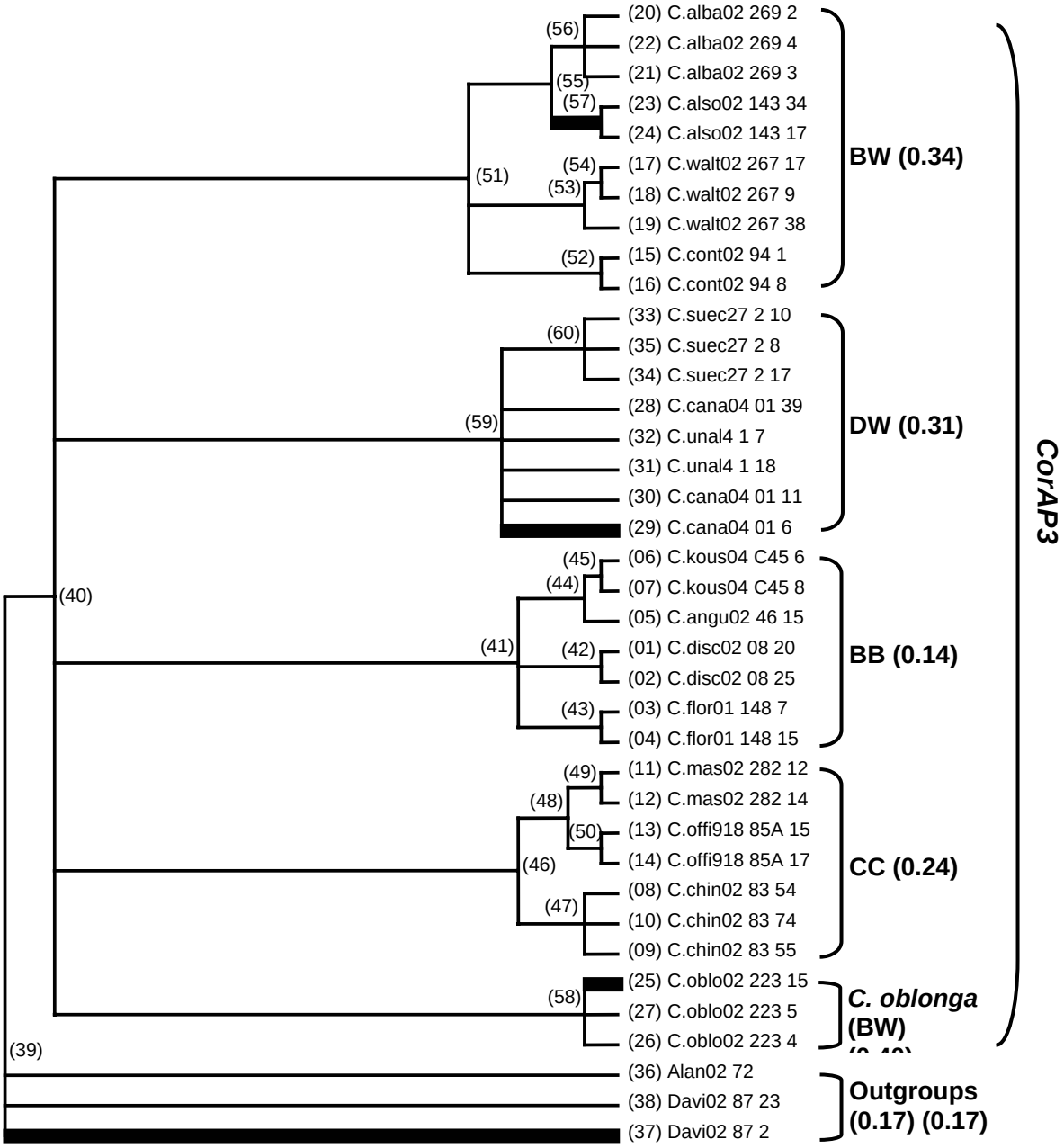

(20) C.alba02 269 2
(56)
(22) C.alba02 269 4
(21) C.alba02 269 3
(55)
(57)
(23) C.also02 143 34
(24) C.also02 143 17
(54)
(17) C.walt02 267 17
(51)
(53)
(18) C.walt02 267 9
(19) C.walt02 267 38
(52)
(15) C.cont02 94 1
(16) C.cont02 94 8
BW (0.34)
(33) C.suec27 2 10
(60)
(35) C.suec27 2 8
(34) C.suec27 2 17
(28) C.cana04 01 39
(59)
(32) C.unal4 1 7
(31) C.unal4 1 18
(30) C.cana04 01 11
(29) C.cana04 01 6
DW (0.31)
(45)
(06) C.kous04 C45 6
(44)
(07) C.kous04 C45 8
(40)
(05) C.angu02 46 15
(41)
(42)
(01) C.disc02 08 20
(02) C.disc02 08 25
(43)
(03) C.flor01 148 7
(04) C.flor01 148 15
BB (0.14)
(49)
(11) C.mas02 282 12
(48)
(12) C.mas02 282 14
(50)
(13) C.offi918 85A 15
(46)
(14) C.offi918 85A 17
(08) C.chin02 83 54
(47)
(10) C.chin02 83 74
(09) C.chin02 83 55
CC (0.24)
(25) C.oblo02 223 15
(58)
(27) C.oblo02 223 5
(26) C.oblo02 223 4
C. oblonga (BW)
(39)
(36) Alan02 72
(38) Davi02 87 23
(37) Davi02 87 2
Outgroups (0.17) (0.17)
CorAP3

Figure 7. *CorAP3* phylogeny applied for PAML analyses which is compatible to coding and introns plus exons analyses (Fig 2 and Fig 3). (1) Allow and estimate different dN/dS values for *CorAP3* and *AP3*-like genes in outgroup genera *Alangium* and *Davidia* (also see Materials and Methods). (2) Allow and estimate different dN/dS values for different morphological subgroups and outgroups. Numbers in parentheses are estimation of dN/dS values for the lineages indicated. Two dN/dS values of outgroups are based on two runs of analyses mentioned above. Thick branches indicate significant increases of dN/dS values detected with ≥2 nonsynonymous substitutions and 0 synonymous substitutions on the branch (see results and discussions). Abbreviations of taxa names are explained in Table 1. DW = dwarf dogwoods, CC = cornelian cherry, BW = blue- or white-fruited dogwoods, and BB = big-bracted dogwoods.

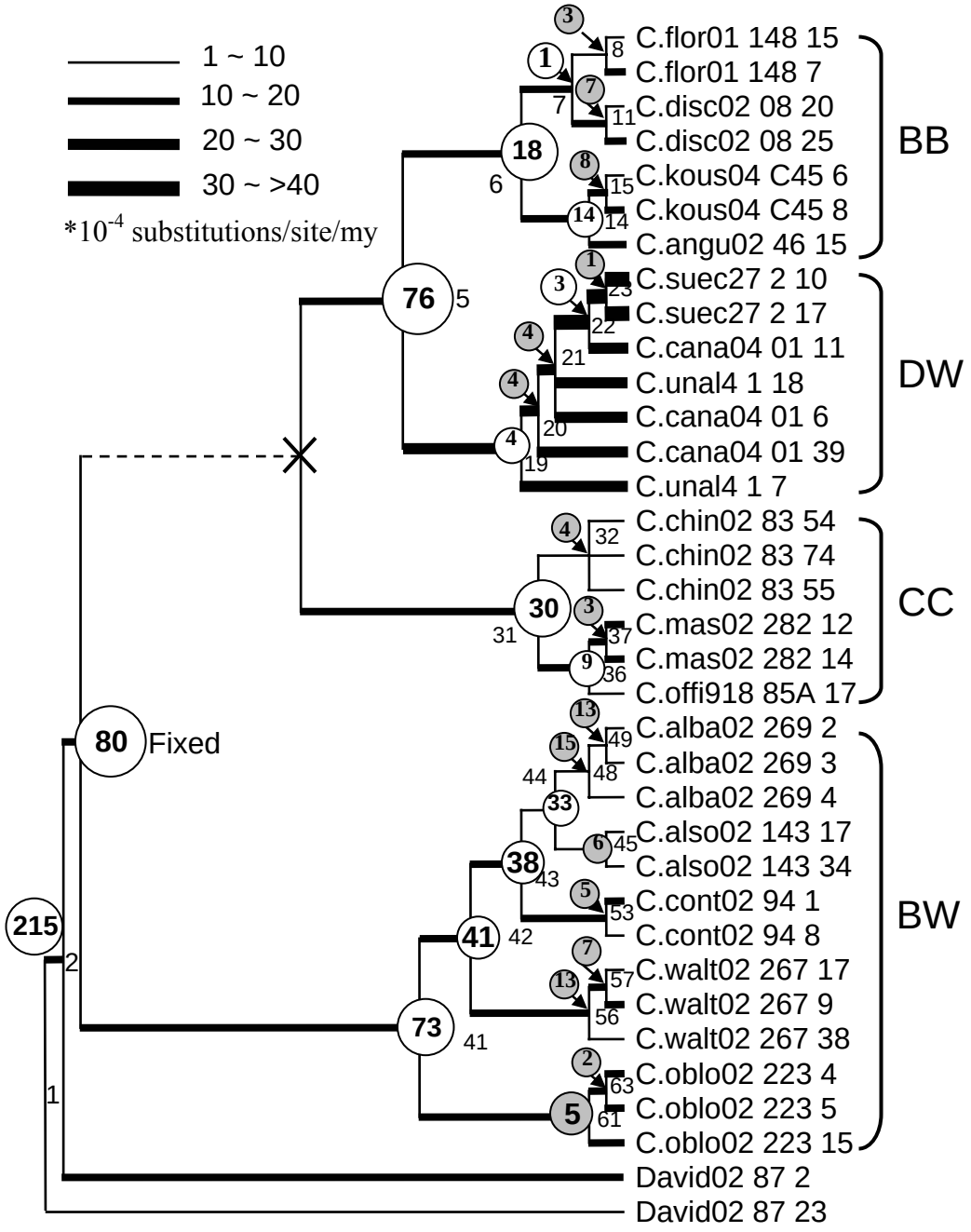

1 ~ 10
10 ~ 20
20 ~ 30
30 ~ >40
*10-4 substitutions/site/my
C.flor01 148 15
C.flor01 148 7
C.disc02 08 20
C.disc02 08 25
C.kous04 C45 6
C.kous04 C45 8
C.angu02 46 15
BB
C.suec27 2 10
C.suec27 2 17
C.cana04 01 11
C.unal4 1 18
C.cana04 01 6
C.cana04 01 39
C.unal4 1 7
DW
C.chin02 83 54
C.chin02 83 74
C.chin02 83 55
C.mas02 282 12
C.mas02 282 14
C.offi918 85A 17
CC
C.alba02 269 2
C.alba02 269 3
C.alba02 269 4
C.also02 143 17
C.also02 143 34
C.cont02 94 1
C.cont02 94 8
C.walt02 267 17
C.walt02 267 9
C.walt02 267 38
C.oblo02 223 4
C.oblo02 223 5
C.oblo02 223 15
BW
David02 87 2
David02 87 23
Fixed

Figure 8. Estimation of divergence times and evolutionary rates of *CorAP3* including both intron and exon regions in r8s 1.71. Topology is constrained to agree with all previous phylogenetic analyses; branch lengths are estimated based on Maximum Likelihood analysis. Node representing the *Cornus* lineage is fixed at 80 mya. Numbers in open circles indicate estimation of divergence time of speciation events; numbers in gray circles indicate gene duplication events. Thickness of branch indicates relative magnitude of substitution rates ($*10^{-4}$ substitutions/site/million year). DW = dwarf dogwoods, CC = cornelian cherry, BW = blue- or white-fruited dogwoods, and BB = big-bracted dogwoods. Numbers near internodes correspond to substitution rates estimated for each branch reported in supplementary materials.

Chapter III

Comparative evolution of two floral homeotic B-class genes *PISTILLATA* and *APETALA3* in the dogwood genus *Cornus* (Cornaceae)

Keywords: B-class genes, MADS-box genes, *PISTILLATA*, *APETALA3*, molecular evolution, gene duplication, selection, coevolution.

Abbreviations

AP3: *APETALA3*

CorAP3: *AP3*-like genes in *Cornus*

PI: *PISTILLATA*

CorPI: *PISTILLATA*-like genes in *Cornus*

CorPI-A, *CorPI*-B: two major copies of *CorPI* in *Cornus*

Abstract:

Discovering the pattern of the floral homeotic gene evolution by comparative study is recognized as an important step toward understanding the genetic basis of morphological evolution. The floral homeotic B-class genes, *APETALA3* (*AP3*) and *PISTILLATA* (*PI*) encode MADS domain-containing transcription factors required to specify petal and stamen identities in *Arabidopsis*. We investigate *AP3*- and *PI*- like gene evolution through genomic DNA in *Cornus*. Gene genealogies and selection pressures between the two B-class genes are compared. Our studies show that ancient gene duplications occurred in *CorPI* genes but were lacking in the *CorAP3* gene lineage during diversification of *Cornus*. Gene duplications within species of *Cornus* are observed for both *CorPI* and *CorAP3*. B-class genes in *Cornus* are generally under purifying selection, but relaxed functional constraints are also observed when compared to B-class genes from other plant groups. We also find relaxation of functional constraints in *CorPI* relative to *CorAP3* lineage, and also in *AP3*- versus in *PI*-like gene lineages of *Alangium* and *Davidia*. Differences in the strength of selection on these paralog loci may be related to maintenance of ancient paralog gene copies in certain lineages. Four positively selected amino acid sites detected in *CorPI* genes indicate that the elevated value of dN/dS in *CorPI* may be a reflection of diversifying selection or positive selection at these sites. Moreover, the positive selection acting on these amino acid sites may be related to functional divergence of the paralog gene copies after gene duplication. For *CorAP3* loci, however, the relaxed functional constraint may be due to neutral evolution. These two nuclear genes provide additional phylogenetically informative variation for resolving phylogenetic relationships at the species level in *Cornus*, but may be difficult to apply widely in phylogenetic studies in other plant groups.

Introduction

Duplication and evolution of floral organ identity genes are considered to play an important role in the origin and evolution of flowers in angiosperms (Theissen et al. 2000; Munster et al. 2001). Most floral organ identity genes belong to the MADS box gene family and are broken into the ABCDE classes (Pelaz et al. 2000; Theissen and Saedler 2001). In *Arabidopsis*, these five classes of genes encode transcription factors that interact to regulate developmental pathways of four series of floral parts. The A class genes consist of *APETALA1* (*AP1*) and *APETALA2* (*AP2*) (*AP2* is not a MADS box gene). The B class genes consist of *APETALA3* (*AP3*) and *PISTILLATA* (*PI*) genes. The C class gene comprises *AGAMOUS* (*AG*), the D class gene consists of *AGAMOUS-LIKE11* (*AGL11*), and the E class genes include *SEPALLATA1* (*SEP1*), *SEPALLATA2* (*SEP2*), and *SEPALLATA3* (*SEP3*). Molecular evolutionary analyses suggest that these different gene classes were established in a relatively short span of evolutionary time and various floral homeotic loci had originated before the appearance of flowering plants (Purugganan et al. 1995). Evidence also indicates that gene duplications occurred within *AP1*, *AP3* and *AG* gene lineages prior to the origin of core eudicots, which led to the hypothesis that duplication and diversification events in these MADS box gene lineages may have been key innovations leading to the radiation of core eudicots (see review in Irish 2003). Detailed studies of the evolutionary history and pattern of MADS box gene diversification in plants may provide useful evidence for evaluating these hypotheses. However, the explicit evolutionary pattern of most floral organ identity genes and their roles in species diversification in most flowering plant lineages remain unknown.

The two B class proteins, *AP3* and *PI,* which determine petal and stamen development contain an N-terminal DNA-binding MADS domain of 58 aa, an interaction I domain, a K region that is predicted to form a coiled-coil domain, and a divergent C terminus (Fig. 1). These proteins are known to form heterodimers to function in regulating the development of petals and stamens (Schwarz-Sommer et al. 1992; Tröbner et al. 1992; Goto and Meyerowitz 1994; Riechmann et al. 1996a; Riechmann et al. 1996b), and are among the most well studied floral organ identity genes from an evolutionary point of view (Irish 2003; Kramer and Jaramillo 2005). Evolutionary studies of mRNA sequences of B-class genes have been conducted on basal angiosperms and lower eudicot lineages (e.g., representatives of Amborellales, Nymphaeales, Austroballeyales, Magnoliales, Piperales, Ranunculales etc. also see Kramer et al. 1998; Kramer and Irish 2000; Kramer et al. 2003; Kim et al. 2004; Stellari et al. 2004; Aoki et al. 2004). *AP3*- and *PI*- like genes are closely related paralogs and gene duplication events producing these two gene lineages predated the origin of angiosperms (Purugganan et al. 1995; Kramer et al. 1998; Aoki et al. 2004; Kim et al. 2004). Subsequent gene duplications and sequence divergence occurred in each of *PI*- and *AP3*-like lineages according to studies in lower eudicots and basal angiosperms (Kramer et al. 1998; Kramer and Irish 2000; Kramer et al. 2003; Kim et al. 2004; Stellari et al. 2004; Aoki et al. 2004). A major gene duplication event is suggested to have occurred in the *AP3* lineage at the base of core eudicots (Kramer et al. 1998). This duplication event is considered to have resulted in two *AP3*-related lineages in core eudicots, termed the *euAP3* lineage that includes *Arabidopsis AP3, Antirrhinum DEF*, and the *TM6* lineage (Kramer et al. 1998; Irish 2003). In contrast, there is no known evidence suggesting a major duplication of *PI*-like gene lineages

during the origin and early diversification of core eudicots (Kramer et al. 1998). Although only a single copy of *AP3*- and *PI*- like genes is found in most of core eudicot species investigated (e.g., *Medicago sativa, Arabidopsis thaliana, Hydrangea macrophylla, Antirrhinum majus, Nicotiana tabacum*), more than two copies of these genes are found in some species of lower eudicots, monocots, and basal angiosperms (e.g., *Ranunculus ficaria, Piper magnificum, Calycanthus floridus, Oryza sativa, Nuphar japonicum, Amborella trichopoda*) (Kramer et al. 1998; Kramer and Irish 2000; Kramer et al. 2003; Kim et al. 2004; Stellari et al. 2004; Aoki et al. 2004). Furthermore, phylogenetic analyses indicate B-class genes duplicated multiple times independently at various phylogenetic levels in lower eudicots and basal angiosperms, i.e., within a species, in a genus, or in the last common ancestor of several families (Kramer et al. 1998; Kramer and Irish 2000; Kramer et al. 2003; Stellari et al. 2004), suggesting a dynamic and stochastic pattern of gene evolution.

Given the functional inter-dependence of two B-class genes, it is expected that changes in these genes especially at the interaction region, the I-domain and partial K-domain (Riechmann et al. 1996a; Krizek and Meyerowitz 1996, Fig. 1) could be correlated or dependent on one another. However, no studies to date have compared sequence changes between these interacting genes in an evolutionary context. In previous two chapters, we examined molecular evolution of B-class genes in dogwood genus *Cornus* (Cornaceae), a higher eudicot lineage by reconstructing the gene evolutionary history and analyzing evolutionary rates and selection pressure (meatrued by dN/dS). In the present study, we compare the rate, pattern, and selection of *AP3* and *PI* DNA sequences. We analyzed the 5' half of genomic DNA sequences of B-class genes containing the I-domain as well as partial

MADS box and K domains, which are regions responsible for their functional specificity (Riechmann et al. 1996a; Krizek and Meyerowitz 1996, Fig. 1). Through analysis of genomic DNA sequences with multiple sampling from each subgroup and its constituent species of *Cornus*, we have obtained detailed genealogies of B-class genes in the genus. We compare the following four aspects to understand evolutionary history, substitution rate, and selection in the evolution of *AP3*- and *PI*- like gene sequences: 1) congruence in species relationships; 2) tempo and frequency of gene duplications; 3) tempo and rate of orthologous sequence divergence; and 4) functional constraints.

Our studies show that ancient gene duplications occurred in *CorPI* genes but were lacking in the *CorAP3* gene lineage during diversification of *Cornus*. Gene duplications within species of *Cornus* are observed in both *CorPI* and *CorAP3* lineages. Functional constraints of B-class genes in *Cornus* are generally relaxed compared to those observed in other plant groups. These relaxed functional constraints are more common in *CorPI* than in *CorAP3*, and also in *AP3*- when compared to the *PI*-like genes of *Alangium* and *Davidia*. The differences in selection on these paralog loci may be related to maintenance of ancient paralog gene copies in the *CorPI* gene lineage in *Cornus*, and also in *AP3*-like gene lineage in *Davidia*. Four positively selected amino acid sites detected on *CorPI* genes indicate that the relaxation of functional constraints in *CorPI* may be due to diversifying selection or positive selection. Moreover, positive selection acting on these sites may be related to gene duplications. However, for *CorAP3* loci, relaxed functional constraints may be due to neutral evolution. Our results support that (1) gene duplication and retention of paralogs may be a general pattern of B-class gene evolution (and MADS box gene evolution) in flowering

plants; (2) gene duplication is an important mechanism which provides raw materials for positive selection or diversifying selection to act on to drive organismal evolution; (3) rates of molecular evolution in the nuclear genome are dynamic and variable among different loci and over time.

Materials and Methods

Comparing evolutionary histories, substitution rates, and selection between *PI*- and *AP3*- like genes

Corresponding regions of *PI*- and *AP3*- like genes are amplified from genomic DNA, which spans the conserved MADS-box domain (exon 1) to the K domain (exon 4) (Fig. 1; also see Materials and Methods in Chapter 1 and 2). We first compare structural differences between two genes by comparing variation in lengths of coding and intron regions. The absolute number and percentage of parsimony informative sites (PIS-parsimonious informative sites, PPIS-percentage of PIS), and those for variable sites (VS-variable sites, PVS-percentage of VS) are also calculated and compared. Furthermore, the numbers of paralog sequences in each species are compared between *CorPI* and *CorAP3*.

Genealogical patterns of *CorPI* and *CorAP3* are compared via phylogenetic analyses. For each gene, analyses of coding region only and of both coding and noncoding regions are carried out. The phylogenetic analyses including only coding regions with taxa from other flowering plants indicate the homology of *Cornus* B-class genes (see taxon sampling in Chapter 1 and 2). Both Neighbor-Joining (NJ) implemented in PAUP* 4.0b10 (Swofford

2002) and Bayesian Metropolis-Hastings coupled Markov chain Monte-Carlo (MCMC) methods in MrBayes 3.1.2 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) are conducted. ModelTest 3.06 (Posada and Crandall, 1998) is first used to select a best-fit model of sequence evolution (details see chapter 1 and 2). The phylogenetic analysis including both coding and noncoding regions is conducted using the same methods. Sequences are aligned using Clustal X (Thompson et al. 1997) and adjusted by eye. Because two *PI* paralogs *CorPI*-A and *CorPI*-B are identified, highly divergent intron sequences of these two paralog genes can not be rationally aligned, only sequences of *CorPI*-A retained in all four subgroups are analyzed. Same phylogenetic analyses are conduncted for *CorPI*-A and *CorAP3* for comparisons.

Information on tempo and rate are estimated using PL method in r8s 1.71 (Sanderson 2002) (detailed see Chapter 1 and 2). We estimate divergence times of taxa lineages and sequence substitution rates using a relaxed clock method, the Penalized Likelihood (PL) method implemented in r8s 1.71 (Sanderson 2002) to see whether divergence of these two genes occurred at similar time periods during diversification of the genus, and whether their sequences evolved at corresponding rates. Both coding and noncoding regions of *CorPI*-A and *CorAP3* are included in analyses. Branch lengths are generated using ML method in PAUP* 4.0b10 (Swofford 2002) based on a fixed topology that reflects relationships of species of the genus (Xiang et al. 2006). Fossil information of *Cornus* is used to place calibration points on the gene genealogy. The node between CC and BB-DW is fixed as 62 mya in the analysis as a minimum age of divergence between the two groups following Xiang et al. (2005) for the *CorPI*-A analysis. However, this node can not be applied for

CorAP3, because there is zero branch length on the node. Therefore, we fix the node between *Cornus* and its outgroup speices as 80 mya, based on the age of the genus estimated using fossils and several other molecular data sets (Xiang et al. unpublished). The smoothing value is applied for these analyses according to the cross-validation analysis imbedded in r8s 1.71 (Sanderson 2002). The estimation of absolute divergence time and substitution rates is done by using the TN algorithm suggested by the program (Sanderson 2002).

Selection constraints are estimated as dN/dS ratios based on pairwise comparison using MEGA 3.1 (Kumar et al. 2004) and codon-based likelihood method in PAML 3.15 (Yang 1997) (detailed see chapter 1 and 2). We measure dN/dS ratios of sequences for *Cornus*, and for different morphological subgroups (BB, DW, CC, and BW), separately. To compare whether there are significant differences of dS, dN, and dN/dS between *CorPI* and *CorAP3* of *Cornus*, *PI*-like genes and *AP3*-like genes in outgroup species *Alangium* and *Davidia*, significant differences of these values based on pairwise comparison method are examined by using two-sample bootstrap resampling test with 10,000 replicates. Six codon-substitution models implemented in Codeml program in PAML 3.15 (Yang 1997, Yang et al. 2000) are also employed to test whether any amino acid sites were under positive selection (methods see Chapter 1 and 2) for both *CorPI* and *CorAP3*.

Results

Sequence structure and variation

The corresponding region of B-class genes is analyzed (Fig 1). The region between two primers is about 1,200 bp and 600 bp for *CorPI* genes and *CorAP3* genes, respectively. The exons have the same sizes, however, lengths of three introns are highly divergent between *CorPI* and *CorAP3* (Fig 2). Intron 1 and intron 2 of *CorPI* are more than two times longer as compared to those of *CorAP3* (Fig 2). The same pattern is also observed in other plant species, e.g., *Antirrhinum* [*GLO* (*PI* homolog): X68831, *DEF* (*AP3* homolog): X62810]. Generally, the numbers of Parsimony Informative Sites (PIS) and Variable Sites (VS) in *CorPI* are larger than those of *CorAP3* in all categories, and the same trend is also observed for Percentage of Parsimony Informative Sites (PPIS) and Percentage of Variable Sites (PVS) (Table 2). The coding region within Cornaceae s. l. is an exception. This reverse may be because *Hydrangea*, a further related species of *Cornus* is included in the *AP3*-like gene analysis, which increases the number of PPIS and PVS. Because the corresponding region covered for *CorPI* is about twice as large as *CorAP3*, more PIS and VS in *CorPI* may be attributed to its longer sequence. However, the PPIS and PVS are also larger in *CorPI* than in *CorAP3*, which might be an indication of increased substitution rates in the *CorPI* gene lineage.

Multiple sequence types of B-class genes within each species are detected. Among 17 identical samples for both *CorPI* and *CorAP3*, 88.2% (15 out of 17) has more sequence types in *CorPI* than in *CorAP3* (Table 3). Phylogenetic analyses show there are two ancient gene duplications occurred in the *CorPI* lineage but are lacking in *CorAP3*, which may have

contributed to a higher number of sequence types in *CorPI*. The difference in polymorphisms between these two genes is likely not due to the experimental bias of sampling, since similar numbers of clones are screened (Table 1). Although sequences of *CorPI* are twice as long as those of *CorAP3*, which may prone to have more Taq errors on the longer sequence and result in a high level of sequence types, this artifact might be insignificant given low rate of Taq error in PCR with short sequences of these two examined sequences. Taq error in PCR is estimated to be $8 * 10^{-6}$ (mutation frequency/bp/duplication) (Cline et al. 1996). For a 1200 bp or 600 bp sequence in this study, there should be less than 1 (actually value 0.58, 0.29) error expected. Therefore, the sequence difference among clones detected can be inferred as arising from alleles, orthologs, or paralogs.

Gene genealogies and substitution rates

Major differences of evolutionary histories in *CorPI* and *CorAP3* according to gene genealogies are several: (1) the *CorAP3* is an ortholog gene lineage without ancient gene duplications detected, while the *CorPI* lineage experienced two ancient gene duplications. One predated before the diversification of *Cornus* to form two paralog genes *CorPI*-A and *CorPI*-B (Chapter 1 – Fig 2). Another is a duplication of *CorPI*-A and had occurred before the diversification of the BW subgroup (Chapter 1 – Fig 2 and 3). (2) The *CorPI*-A genealogy resolved the relationships well among four morphological subgroups (BB, DW, CC, and BW) (Chapter 1 – Fig 2 and 3), while *CorAP3* genealogy shows a polytomy among them (Chapter 2 – Fig 2 and 3). (3) The conflict resolution of relationships of *C. unalaschkensis*, the 4n species with its putative parent species *C. canadensis* and *C. suecica*

based on *CorPI* and *CorAP3*. Phylogenetic analyses of *CorPI* sequences show close relationship between *C. unalaschkensis* and *C. suecica*. However, this 4n species embedded within sequences of *C. canadensis* in *CorAP3* analyses. Despite these differences, the gene genealogies of *CorPI* and *CorAP3* also share some similarities: (1) multiple sequence types are detected for most species sampled (also see above), suggesting gene duplication took place within species. (2) The species relationships within each major subgroup are well recovered indicating that the orthologs of both nuclear genes diverged along with speciation events. In addition, two divergent copies of *AP3*-like gene in *Davidia* are found (see detailed statistics in Chapter 2). Since these two sequences are highly divergent, it may be an indication that the gene duplication forming these two paralog genes is ancient, followed by the loss of one gene copy in ancestors of *Alangium* and *Cornus*. However, we could not completely rule out the possibility of accelerated substitution rate change after the *AP3*-like gene duplication within *Davidia* (see Chapter 2).

Nucleotide substitution rates of *CorPI* and *CorAP3* are unequal and heterogeneous in *Cornus* lineages. Cross-validation analyses indicate that the best smoothing parameters are 1 and 3 for *CorPI*-A and *CorAP3*, respectively, which imply great deviation from clocklike rate changes (smoothing parameter towards 1000) of both genes. The results indicate that DW has the highest substitution rate (substitution/site/my) among all other subgroups of *Cornus*, with a value of $3.9 *10^{-3} \sim 4.2 *10^{-3}$ in *CorPI*-A, and $2.2 *10^{-3} \sim 4.0 *10^{-3}$ in *CorAP3*. The substitution rates observed in DW are generally 1.5 to 4 times greater than those in other subgroups. The results of 2-sample bootstrap resampling tests demonstrate that substitution rates of *CorPI*-A is significantly higher than those of *CorAP3* in DW ($P = 0.001$),

and in BB (P = 0.04). In contrast, substitution rates of *CorAP3* are significantly greater than rates of *CorPI*-A in the CC (P=0.008). Substitution rates between *CorPI*-A and *CorAP3* in BW are not significantly different (P=0.17).

Although heterogeneous substitution rates of the gene evolution at different time and along different morphological lineages are considered, the estimates of lineage divergence time based on *CorPI*-A and *CorAP3* are still distinct at deep nodes (nodes 1, 2, and 8, Fig 3) which represent early speciation events of the genus leading to major morphological subgroups. At nodes 1, 2 and 3 *CorAP3* diverged prior to the late Cretaceous, and are 18, 28 and 21 million years older than the divergence of *CorPI* (Fig 3). One explanation for these discrepancies is that evolutionary histories of these two nuclear genes are intrinsically different in tempo of divergence. The other explanation is that limited informative sites of *CorAP3* gene result in lacking branch lengths to support those deep nodes (see above about the *CorAP3* gene genealogy). In either case, it is the hard polytomy in the *CorAP3* phylogeny which may cause the bias of time estimation at those basal nodes. Therefore, to distinguish the real evolutionary pattern of *CorAP3* from the sampling bias, investigation of the full length of the gene is required. For all other nodes, the estimates of divergence time are comparable between *PI* and *AP3*. Substitution rates of *CorPI* genes on braches leading to nodes 1 - 4 and 6 are clearly higher than rates of *CorAP3*. Interestingly, these branches lead to two clades BB and DW that had evolved petaloid bracts.

Functional constraints and selection

The dN/dS ratios are substantially different between *AP3*- and *PI*- like genes in *Cornus* and outgroup taxa. The mean dN/dS ratio of *CorAP3* gene lineage based on pairwise comparison analysis (0.33 ± 0.16) is significant smaller than those (0.49 ± 0.38) in *CorPI* (P = 0.0001). Contrarily, the mean dN/dS value of *AP3*-like gene lineage of *Davidia* and *Alangium* (0.28 ± 0.03) is higher than those values (0.13 ± 0.10) for *PI*-like gene lineage, but the difference is not proven statistically significant (P=0.09). Noticeably, in both cases, the lineages which experienced more gene duplication events and kept more paralogous gene copies, i.e., *CorPI* genes in *Cornus* and *AP3*-like genes in outgroups, tend to have greater dN/dS ratios as compared to their counterparts. The nonsynonymous substitution rate (dN) is significantly less than the synonymous substitution rate (dS) calculated for all categories, i.e., *CorPI*, *CorAP3*, *PI*-like genes in outgroups, and *AP3*-like genes in outgroups, suggesting the substitution rate on nonsynonymous sites is generally strongly constrained. The 2-sample bootstrap resampling tests indicate that dS and dN of *CorAP3* (0.02 ~ 0.39, ave. 0.18; 0 ~ 0.13, ave. 0.06) are significantly greater than dS and dN of *CorPI* (0 ~ 0.47, ave. 0.13; 0 ~ 0.14, ave. 0.05) (P = 0.0001 in both cases), although the dN/dS ratio in *CorPI* is greater compared to *CorAP3* (see above). For the outgroup species, dN is significant higher in *AP3*- (0.11 ~ 0.22, ave. 0.17) than in *PI*-like genes (0.01 ~ 0.05, ave. 0.03) (P = 0.01), but no significant difference in dS between *AP3*- (0.37 ~ 0.77, ave. 0.63) and *PI* -like genes (0.02 ~ 0.43, ave. 0.27) (P = 0.09).

The strength of selection values (dN/dS ratio) estimates in *CorPI* and *CorAP3* based on both pairwise comparison using MEGA and codon-based likelihood method using PAML are

congruent, although the dN/dS values estimated by PAML are generally less than those by MEGA, except for *CorPI*-B copy (Table 4). Intriguingly, all the corresponding dN/dS ratios observed in *CorPI*-A are greater than those in *CorAP3* for all categories, suggesting functional constraints are more relaxed in *CorPI*-A compared to *CorAP3*. The greater values of dN/dS in *CorPI*-A compared to those in *CorAP3* are statistically significant (P = 0.0001) based on 2-sample bootstrap resampling tests. The increased dN/dS values of *CorPI*-A compared to *CorAP3* within each subgroup are also significant, which is in DW (P = 0.0001), in BB (P = 0.002), and in BW (P = 0.0001), except in CC (P = 0.13) which is not statistically significant.

Furthermore, for *CorPI* genes, 94.6% of dN/dS values are less than 1 based on pairwise comparison results, and 5.4% of dN/dS values are between 1 and 4. For *CorAP3*, 98.3% of dN/dS values are less than 1, and 1.7% are between 1 and 1.5. The higher percentage of dN/dS ratios over 1 in *CorPI* is consistent with the relaxation of functional constraints or presence of positive selection (see below) in *CorPI* compared to *CorAP3*. The dN/dS ratios of each branch on the phylogeny estimated based on codon-based likelihood method indicate the similar trend. For *CorAP3* gene, 16 out of total 59 branches (27%) have a dN/dS value over 1, and the maximum dN/dS value is 257.32 and the minimum dN/dS value is 0.0001. For *CorPI* gene, 45 out of 133 braches (34%) have a dN/dS ratio over 1, which demonstrates more branches have large dN/dS ratio compared to *CorAP3*, and the maximum dN/dS value is 999, while the minimum dN/dS value is 0.0001.

The positive selection is detected at four amino acid sites in *CorPI*, three out of these four sites are located in the I-domain. The amino acid site I_4 is identified to be under positive selection at the 99% level by all models that allow for positive selection. Sites M_{54}, I_6, and I_{11} are also suggested under positive selection at the 95% level by models M2, M3 and M8 (see Table 4, in Chapter 1). However, no positive selection on any amino acid sites are found for *CorAP3*, and none of the models that allows for the presence of positively selected sites in *CorAP3* is significantly better than alternative models (see Table 2, in Chapter 2). Only the model M3 has a significantly higher likelihood value than the model M0, and none of M2 and M8 models is a better fit for our data than their counterparts M1 and M7 which do not allow for positive selection (Table 3, in chapter 2), suggesting no positive selection acted on the gene and all amino acid sites evolved generally homogenously.

Discussion

Gene duplications accompanied B-class gene evolution in flowering plants

Gene duplications in *AP3* and *PI* lineages occurred at various phylogenetic levels during the origin and diversification of flowering plants (see review Irish 2003). Two distinct *AP3* lineages, eu*AP3* and *TM6* originated through an ancient gene duplication event before the diversification of higher eudicots (Kramer et al. 1998). The *euAP3* copy retains the function of *paleoAP3* of identifying the petal and stamen development, whereas the function of *TM6* gene lineage is still unclear (Irish 2003). None of gene duplication events, however, found in the *PI* lineage appears to date to the base of higher eudicots (Krammer et al. 1998), except for an ancient duplication event in the *PI* lineage that predated the split between Lauraceae and

Calycanthaceae (Stellari et al. 2004). Different paralog pairs have been identified in the *PI* lineage in some species, majorly lower eudicots and basal angiosperms, e.g., *Petunia hybrida, Ranunculus bulbosus, Papaver nudicaule, Piper magnificum, Oryza sativa* and *Nuphar japonicum*, and phylogenetic analyses suggest that these paralogous gene copies are the products of independent duplication events which occurred within a species or its genus (Kramer et al. 1998; Kramer et al. 2003; Aoki et al. 2004; Stellari et al. 2004; Kim et al. 2004). In two of these species, *P. hybrida* and *O. sativa*, there is some experimental evidence that demonstrate the functional divergence of paralog gene copies (Angenent et al. 1993; Chung et al. 1995). Although polyploidy is a common cause for the presence of paralogs within plant genomes, none of these species is known to be polyploids (Bennett and Smith 1976, 1991; Bennett and Leitch 1995), suggesting independent origin by gene duplication within these plant species. In our study, we detect duplications of *PI*-like genes at various phylogenetic levels in *Cornus*, with one predating the diversification of *Cornus* to form two paralog genes *CorPI*-A and *CorPI*-B (Chapter 1 – Fig 2), one before the diversification of a clade in the BW subgroup (Chapter 1 – Fig 2 and 3), and we find evidnece of many independent duplication events within species (Fig 2 and 3 in Chapter 1). Therefore, our findings in *Cornus*, a group of core eudicots, agree with previous studies suggesting dynamic gene duplications and a trend toward maintenance of multiple copies of *PI* homologs in lower eudicots and basal angiosperms (Kramer et al. 1998; Kramer et al. 2003; Aoki et al. 2004; Stellari et al. 2004; Kim et al. 2004).

Like *PI* gene lineage, multiple gene copies of *AP3*-like gene are detected mostly in lower eudicots and basal angiosperms (Kramer et al. 1998; Kramer et al. 2003; Aoki et al.

2004; Stellari et al. 2004; Kim et al. 2004). Taxon-specific duplication events are found in *AP3* lineages, e.g., *Brassica oleracea, Rumex acetosa, Pachysandra terminalis,* and *Papaver nudicaule*. Evidence for a functionally divergent *AP3* paralog is seen in the *Medicago sativa* gene, *NMH7* (Heard and Dunn 1995). The sequence of *NMH7* is highly diverged from sequences of other *AP3* lineage members, and may reflect the fact that the function of this paralog has also diverged considerably, being known to be involved in mediating root nodulation (Heard and Dunn 1995). In *Cornus*, there is no ancient gene duplications detected in *AP3* gene lineage. We find divergent copies of *AP3*-like genes in *Davidia*, the close relative of *Cornus*. Interestingly, one of two *AP3* copies of *Davidia* having 12 nonsynonymous while zero synonymous changes are detected, suggesting strong positive selection and functional divergence on this gene copy. Results reported here for *Cornus AP3* gene lineage show that gene duplications occurred independently within most of *Cornus* species (Fig 2 and 3 in Chapter 2), which is congruent with findings in lower eudicots and basal angiosperms.

Taken together, phylogenetic analyses indicate that B-class gene duplications took place at various phylogenetic times in *Cornus*. This result agrees with recent understandings about B-class gene evolution in basal angiosperms and lower eudicots, suggesting a dynamic and stochastic pattern of B-class gene evolution (Stellari et al. 2004, Kramer et al. 1998; Kramer and Irish 2000; Kramer et al. 2003).

Does selection associate with gene duplications?

It has been demonstrated theoretically that positive selection is required in the early stage of evolution of a new function from a duplicate gene (Li 1997). A recent empirical study indicated that selection, rather than drift, plays a role in the establishment of duplicate gene loci, as evidenced by finding the key role of positive selection in preserving some gene copies (Moore and Purugganan 2003). This study also found that positive selection indeed can act at a very early stage of the gene duplication process (Moore and Purugganan 2003). It has been speculated that gene duplication may contribute greatly to positive selection observed during regulatory gene evolution (Van de peer et al. 2001; Fares et al. 2003; Martinez-Castilla and Alvarez-Buylla 2003). However, subsequent rapid evolution within a single orthologous copy of regulatory genes may be primarily due to neutral evolution (Chang et al. 2005). Our results indicate a strong correlation between more relaxed functional constraints or diversifying selection with gene duplications (see results). We find dN/dS ratios of *CorPI* are significantly larger than those of *CorAP3* in *Cornus*, and the dN/dS of *AP3*-like genes is larger than the dN/dS of *PI*-like genes in *Davidia* and *Alangium*. These results suggest that functional constraints in *CorPI* genes and *AP3*-like genes are more relaxed than their counterparts *CorAP3* and *PI*-like genes, which might correlate with gene duplications observed in *CorPI* and *AP3*-like lineages. *CorPI* genes are detected having four amino acid sites under strong positive selection suggesting positive selection acted on the gene, but no positively selected amino acid sites are detected for *CorAP3* genes (see PAML results). General trends of larger dN/dS ratios in *CorPI* than in *CorAP3* are further demonstrated for all subgroup comparisons (Table 4). A recent study observed the less constrained *PI*-like gene evolution (0.19 ± 0.02) compared to other floral homeotic genes

including the *AP3*-like gene (0.12 ± 0.03) but only based on a broad scattered sampling (Purugganan et al. 1995). Whether this accelerated dN/dS ratio of *PI*-like gene clade is also associated with gene duplication in this analysis is not clear due to the lack of knowledge of gene evolutionary patterns based on the sparse species sampling. The phylogenetic analysis of the entire plant MADS box gene family has shown that *AP3* and *PI* lineages are products of a gene duplication event which makes them more closely related to each other than to any of the other MADS-box genes (Doyle 1994; Purugganan et al. 1995; Purugganan 1997; Theissen et al. 1996). *PI* and *AP3* genes have been known to function as protein heterodimers to specify the petal and stamen development in flowering plants. Because of the similar function between these paralog genes, functional constraints acting on them should be alike. Therefore, I speculate that the relaxation of functional constraints observed may be due to the maintenance of paralog genes in *CorPI* of *Cornus* and *AP3*-like genes of *Davida*. Although functional constraints on both *CorPI* and *CorAP3* are relaxed compared to other flowering plants (see below), the relaxation of functional constraints on *CorPI* may be due to positive selection acted on certain lineage and certain amino acids of the gene, however, for *CorAP3* it is possibly due to neutral evolution. We don't know whether this rapid change in *CorAP3* may also be partly driven by coevolutionary mechanisms due to interacting with *CorPI*.

Changes in the encoded proteins of a homeotic gene are thought to play a prominent role in the evolution of new morphologies (Lamb and Irish 2003). The functional specificity and the interaction of B-class genes, *PI* and *AP3,* are determined by I- and K- domains (Krizek and Meyerowitz 1996). We find four amino acid sites are under strong positive selection on *CorPI* genes. Interestingly, three located on the I-domain, which is a region between the

MADS domain and K-domain (Figure 1) and plays an important role in dimerization specificity (Riechmann et al. 1996a). The I-domain region is known to be only relatively weakly conserved compared to the MADS box and K domains. It is generally excluded from the analysis for distant related plants (Purugganan et al. 1995) and for highly divergent gene members, i.e., the complete MADS box gene family (Martinez-castilla and Alvarez-buylla 2003) to avoid detecting false-positive selection at amino acid sites. Considering the critical function of the I-domain in forming heterodimers between *PI* and *AP3* in order to specify the petal and stamen development (Riechmann et al. 1996a; Krizek and Meyerowitz 1996), it is necessary to understand the molecular evolution on the I-region. Since the close relativeness of species sampled for our study, there is no uncertainty about the sequence alignment and no saturation of nucleotide substitution on the I-domain. Therefore, these positively selected amino acids detected in the I-domain of *CorPI* genes, which also involved changes of polarity and charges of amio acids, may have brought profound changes or modification to its regulatory function. The combination of molecular evolution and functional studies is necessary to reveal the significance of these positively selected amino acid sites in the functional divergence after gene duplication and during the diversification of *Cornus*.

The functional constraints on the B-class genes in *Cornus* are identified relaxed compared to *PI*- and *AP3*-like genes in other flowering plants. dN/dS ratios of *CorPI* (0.49 ± 0.38) and *CorAP3* (0.33 ± 0.16) in *Cornus* are generally larger than the mean dN/dS value (0.14) detected for other plant nuclear loci (Purugganan 1998) and 0.19 ± 0.02 for *PI*-like genes and 0.12 ± 0.03 for *AP3*-like genes found in other plants (Purugganan et al. 1995), which suggests that relaxation in functional constraints of these regulatory genes may have

adaptive significance during the diversification of this genus. Molecular evolutionary analyses also reveal that there are appreciable differences in the substitution rates between different domains of plant MADS-box genes (Purugganan et al. 1995). In this study, only the 5' region of B-class genes is investigated, which covered partial MADS-box, complete I- and one third K- domains. Because the I-domain together with C-domain are known to be more variable than MADS box and K- domain (Purugganan et al. 1995), the increased dN/dS ratios observed in this study for B-class genes in *Cornus* may be due to the bias of sampling in accelerated evolved regions. The molecular evolution study of MADS-box genes in other flowering plants demonstrates that the selection pressure on K- and C- domains is more relaxed than on other MADS-box regions (0.04 ± 0.01 and 0.10 ± 0.01, for *AP3* and *PI*, respectively). The dN/dS values of the K domain are 0.17 ± 0.05 and 0.27 ± 0.03 in *AP3* and *PI*, respectively, and the dN/dS values of the C domain are 0.18 ± 0.03 and 0.23 ± 0.03 in *AP3* and *PI*, respectively (Purugganan et al. 1995). Compared to these values, the substantial relaxation of functional constraints in *CorPI* (0.49 ± 0.38) and *CorAP3* (0.33 ± 0.16) observed in this study might not be a bias, however to obtain a complete sequence of these genes is necessary to confirm this conclusion. Moreover, the fact that petals and stamens display enormous morphological plasticity within angiosperms may be directly reflected as changes in B-class genes. These predictions are supported by a substitution rate analysis, which shows that the *AP3*/*PI* group is evolving 20–40% faster than all of the other plant MADS-box genes (Purugganan 1997). I also observe relatively accelerated dN/dS ratios in BB and DW subgroups with petaloid bracts morphology in *Cornus*, but there is no evidence of a correlation between B-class gene evolution and evolving petaloid bracts.

Nuclear genes as phylogenetic markers – benefits and challenges

Recent studies indicate that single- or low- copy nuclear genes in plants are a rich source of phylogenetic information at different levels, especially for interspecific relationships which are poorly resolved in plants (Sang 2002; Hughes et al. 2006). Other studies discover that regulatory genes evolve more rapidly than structural genes. Therefore, these genes may be good candidates for investigating phylogenetic relationships in plants. According to the statistics of *PI*- and *AP3*- like genes in *Cornus*, *CorPI* has 30.39% PIS, and *CorAP3* has 22.22% PIS. A previous study on the Myc-like anthocyanin regulatory gene in *Cornus* finds 20.44% PIS of coding regions (Fan et al. 2004). These values of nuclear loci are much higher than those in chloroplast genes (2.77%) and 26S rDNA (4.05%) of *Cornus* (Fan et al. 2004). Furthermore, when noncoding regions are included, the PIS of *CorPI* (42.02%) and *CorAP3* (37.00%) are more than two times larger compared to those values (15%) of chloroplast genes in Cornaceae (Xiang et al. 1998). The relationships at species levels of *Cornus* are fully resolved based on these two B-class genes (Fig 2, 3; chapter 1 and 2).

There are lots of challenges to explore nuclear markers. Those challenges include the extra experiments to separate multiple copies and to identify ortholog gene copies for the reconstruction of organismal relationships. If the purpose of study is getting more phylogenetically informative sites for the phylogenetic reconstruction of organisms, it is ideal to obtain copy-specific PCR primers to directly amplify ortholog genes. However, in many cases, it is difficult to design primers for ortholog genes due to highly conserved exon regions among paralogs, but highly divergent noncoding regions within orthologs. In addition, gene duplications and random gene losses are prevalent in nuclear genes, therefore

extreme caution needs to be paid to infer the relationships of organism if based on only a single nuclear gene. Over all, mining low copy nuclear genes for the phylogenetic reconstruction is generally rewarding, especially for resloving species and interspecies relationships.

It is known that the evolutionary history of different nuclear loci may evolve differently. This difference may be due to the dynamic genome change or the selection pressure acting on these gene loci. In this study, I find the evolutionary rates and patterns of *CorPI* and *CorAP3* are not accordant. First, the substitution rates of *CorPI* including both coding and noncoding regions are higher than those of *CorAP3*. It is evidenced by showing that rates of some lineages of *Cornus* are accelerated in *CorPI*-A than those in *CorAP3* (see r8s results). The noncoding regions between *Cornus* and outgroup genera *Alangium* and *Davidia* are alignable for *CorAP3* but not for *CorPI*, suggesting accelerated substitution rates of *PI*-like gene lineage. Second, there is lacking a resolution of relationships among subgroups in *Cornus* based on *CorAP3*, although well resolved for *CorPI*-A. This may indicate that the dynamics of genome evolution is stochastic. The rate discrepancy for genes with similar function, e.g., *PISTILLATA* and *APETALA3*, may be correlated to their specific evolutionary history at different nuclear loci. Third, the hypothesized allotetraploid species *C. unalaschkensis* (4n = 44) shows closely related to the diploid species *C. suecica* (2n = 22) in *CorPI*-A analysis, but this 4n speices reveals close relationships to another diploid species *C. canadensis* (2n = 22) in *CorAP3* analysis. Discrepancy of phylogenetic position of *C. unalaschkensis* based on different nuclear loci can be explained as differential losses of the redundant gene loci inherited from its diploid parent species. Together with the observation of the anthocyanin

gene which shows that *C. unalaschkensis* and *C. canadensis* are more closely related (Fan et al. 2006), my observations agree with recent understandings about the importance of differential losses of redundant gene copies following hybridization during polyploid plant evolution (Adams et al. 2003). To interprete the evolutionary history of organisms, combining data from multiple genes and different genomes (nuclear, chloroplast, and mitochondria) is necessary to reconstruct a dependable evolutionary history of organisms.

The molecular evolutionary rates are heterogeneous among different organism lineages (lineage effects, see chapter 1) and among distinct gene loci. I find that, in subgroups DW and BB, the substitution rates of *CorPI*-A are significantly larger than those observed in *CorAP3*, which indicate *CorPI*-A might generally evolve faster than *CorAP3* in *Cornus*. However, this trend is not observed for CC and BW subgroups. Moreover, we observe an increase of rates on those branches leading to DW and BB clades (Fig 4). The substitution rates of DW are remarkably increased compared to other subgroup species in both cases of *CorPI* and *CorAP3* indicating short generation time may be a key factor to accelerate evolutionary rates in this herbaceous group as discussed in the previous chapters.

Acknowledgement

We thank Drs. Jer-Ming Hu and Michael Frohlic who provided primers and mRNA sequence data of *C. florida* and *C.alba* for this study. We are also grateful to the NSF supported DEEPTIME program (funded to D. E. Soltis DEB-0090283) for the travel support to workshops on divergence time dating and application of fossil data. We thank M. D.

Purugganan, B. M. Wiegmann, and N. S. Allen for the insightful comments on the early version of the manuscript. We thank D. Thomas and K. M. Ding for helping with collecting data. This study is supported by National Science Foundation grants to Q-Y (J.) X (NSF-DEB 0129069 and DEB 0444125).

References

ADAMS, K. L., R. CRONN, R. PERCIFIELD, AND J. F. WENDEL. 2003. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. Proceedings of the National Academy of Sciences of the United States of America 100: 4649-4654.

ANGENENT, G. C., J. FRANKEN, M. BUSSCHER, L. COLOMBO, AND A. J. VAN TUNEN. 1993. Petal and stamen formation in petunia is regulated by the homeotic gene *fbp1*. Plant Journal 4: 101-112.

AOKI, S., K. UEHARA, M. IMAFUKU, M. HASEBE, AND M. ITO. 2004. Phylogeny and divergence of basal angiosperms inferred from *APETALA3*- and *PISTILLATA*-like MADS-box genes. Journal of Plant Research 117: 229-244.

BENNETT, M. D. AND J. B. SMITH. 1976. Nuclear DNA amounts in angiosperms. Philosophical Transactions of the Royal Society of London B 274: 227-274.

-----. 1991. Nuclear DNA amounts in angiosperms. Philosophical Transactions of the Royal Society of London B 334: 309-345.

BENNETT, M. D. AND I. J. LEITCH. 1995. Nuclear DNA amounts in angioperms. Annals of Botany 76: 113-176.

CHANG, S. M., Y. LU, AND M. D. RAUSHER. 2005. Neutral evolution of the nonbinding region of the anthocyanin regulatory gene *Ipmyb1* in *Ipomoea*. Genetics 170: 1967-1978.

CHUNG, Y. Y., S. R. KIM, H. G. KANG, Y. S. NOH, M. C. PARK, D. FINKEL, AND G. H. AN. 1995. Characterization of two rice MADS box genes homologous to *Globosa*. Plant Science 109: 45-56.

CLINE, J., J. C. BRAMAN, AND H. H. HOGREFE. 1996. PCR fidelity of *pfu* DNA polymerase and other thermostable DNA polymerases. Nucleic Acids Research 24: 3546-3551.

DOYLE JJ. 1994. Evolution of a plant homeotic multigene family - toward connecting molecular systematics and molecular developmental genetics. Systematic Biology 43: 307-328.

FAN, C., M. D. PURUGGANAN, D. T. THOMAS, B. M. WIEGMANN, AND J. Q. XIANG. 2004. Heterogeneous evolution of the *Myc*-like Anthocyanin regulatory gene and its phylogenetic utility in *Cornus* L. (Cornaceae). Molecular Phylogenetics and Evolution 33: 580-594.

FAN, C., Q. Y. XIANG, D. L. REMINGTON, M. D. PURUGGANAN, AND B. M. WIEGMANN. 2006. Evolutionary patterns in the *antR-Cor* gene in the dwarf dogwood complex (*Cornus*, Cornaceae). Genetica. DOI 10.1007/s 10709-006-0016-3

FARES, M. A., D. BEZEMER, A. MOYA, AND I. MARIN. 2003. Selection on coding regions determined *Hox7* genes evolution. Molecular Biology and Evolution 20: 2104-2112.

GOTO, K. AND E. M. MEYEROWITZ. 1994. Function and regulation of the *Arabidopsis* floral homeotic gene *PISTILLATA*. Genes & Development 8: 1548-1560.

HEARD, J. AND K. DUNN. 1995. Symbiotic induction of a MADS-box gene during development of alfalfa root nodules. Proceedings of the National Academy of Sciences of the United States of America 92: 5273-5277.

HUELSENBECK, J. P. AND F. RONQUIST. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17: 754-755.

HUGHES, C. E., R. J. EASTWOOD, AND C. D. BAILEY. 2006. From famine to feast? Selecting nuclear DNA sequence loci for plant species-level phylogeny reconstruction. Philosophical Transactions of the Royal Society of London B 361: 211-225.

IRISH, V. F. 2003. The evolution of floral homeotic gene function. Bioessays 25: 637-646.

KIM, S. T., M. J. YOO, V. A. ALBERT, J. S. FARRIS, P. S. SOLTIS, AND D. E. SOLTIS. 2004. Phylogeny and diversification of B-function MADS-box genes in angiosperms:

Evolutionary and functional implications of a 260-million-year-old duplication. American Journal of Botany 91: 2102-2118.

KRAMER, E. M., V. S. DI STILIO, AND P. M. SCHLUTER. 2003. Complex patterns of gene duplication in the *APETALA3* and *PISTILLATA* lineages of the Ranunculaceae. International Journal of Plant Sciences 164: 1-11.

KRAMER, E. M., R. L. DORIT, AND V. F. IRISH. 1998. Molecular evolution of genes controlling petal and stamen development: duplication and divergence within the *APETALA3* and *PISTILLATA* MADS-box gene lineages. Genetics 149: 765-783.

KRAMER, E. M. AND V. F. IRISH. 2000. Evolution of the petal and stamen developmental programs: Evidence from comparative studies of the lower eudicots and basal angiosperms. International Journal of Plant Sciences 161: S29-S40.

KRAMER, E. M. AND M. A. JARAMILLO. 2005. Genetic basis for innovations in floral organ identity. Journal of Experimental Zoology (Molecular and Developmental Evolution) 304B: 526-535.

KRIZEK, B. A. AND E. M. MEYEROWITZ. 1996. Mapping the protein regions responsible for the functional specificities of the *Arabidopsis* MADS domain organ-identity proteins. Proceedings of the National Academy of Sciences of the United States of America 93: 4063-4070.

KUMAR, S., K. TAMURA, AND M. NEI. 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Briefings in Bioinformatics 5: 150-163.

LAMB, R. S. AND V. F. IRISH. 2003. Functional divergence within the *APETALA3*/*PISTILLATA* floral homeotic gene lineages. Proceedings of the National Academy of Sciences of the United States of America 100: 6558-6563.

LI, W. H. 1997. Molecular evolution. Sinauer Associates, Sunderland, Mass.

MARTINEZ-CASTILLA, L. P. AND E. R. ALVAREZ-BUYLLA. 2003. Adaptive evolution in the *Arabidopsis* MADS-box gene family inferred from its complete resolved phylogeny. Proceedings of the National Academy of Sciences of the United States of America 100: 13407-13412.

MOORE, R. C. AND M. D. PURUGGANAN. 2003. The early stages of duplicate gene evolution. Proceedings of the National Academy of Sciences of the United States of America 100: 15682-15687.

MUNSTER, T., L. U. WINGEN, W. FAIGL, S. WERTH, H. SAEDLER, AND G. THEISSEN. 2001. Characterization of three *GLOBOSA*-like MADS-box genes from maize: evidence for ancient paralogy in one class of floral homeotic B-function genes of grasses. Gene 262: 1-13.

PELAZ, S., G. S. DITTA, E. BAUMANN, E. WISMAN, AND M. F. YANOFSKY. 2000. B and C floral organ identity functions require *SEPALLATA* MADS-box genes. Nature 405: 200-203.

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

PURUGGANAN, M. D. 1997. The MADS-box floral homeotic gene lineages predate the origin of seed plants: Phylogenetic and molecular clock estimates. Journal of Molecular Evolution 45: 392-396.

-----. 1998. The molecular evolution of development. Bioessays 20: 700-711.

PURUGGANAN, M. D., S. D. ROUNSLEY, R. J. SCHMIDT, AND M. F. YANOFSKY. 1995. Molecular evolution of flower development: diversification of the plant MADS-box regulatory gene family. Genetics 140: 345-356.

RIECHMANN, J. L., B. A. KRIZEK, AND E. M. MEYEROWITZ. 1996a. Dimerization specificity of *Arabidopsis* MADS domain homeotic proteins *APETALA1*, *APETALA3*, *PISTILLATA*, and *AGAMOUS*. Proceedings of the National Academy of Sciences of the United States of America 93: 4793-4798.

RIECHMANN, J. L., M. WANG, AND E. M. MEYEROWITZ. 1996b. DNA-binding properties of *Arabidopsis* MADS domain homeotic proteins *APETALA1*, *APETALA3*, *PISTILLATA* and *AGAMOUS*. Nucleic Acids Research 24: 3134-3141.

RONQUIST, F. AND J. P. HUELSENBECK. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572-1574.

SANDERSON, M. J. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Molecular Biology and Evolution 19: 101-109.

SANG, T. 2002. Utility of low-copy nuclear gene sequences in plant phylogenetics. Critical Reviews in Biochemistry and Molecular Biology 37: 121-147.

SCHWARZ-SOMMER, Z., I. HUE, P. HUIJSER, P. J. FLOR, R. HANSEN, F. TETENS, W. E. LONNIG, H. SAEDLER, AND H. SOMMER. 1992. Characterization of the *Antirrhinum* floral homeotic MADS-box gene *deficiens*: evidence for DNA binding and autoregulation of its persistent expression throughout flower development. EMBO Joural 11: 251-263.

STELLARI, G. M., M. A. JARAMILLO, AND E. M. KRAMER. 2004. Evolution of the *APETALA3* and *PISTILLATA* lineages of MADS-box-containing genes in the basal angiosperms. Molecular Biology and Evolution 21: 506-519.

SWOFFORD, D. L. 2002. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4.10b. Sinauer Associates, Sunderland, Mass.

THEISSEN, G. 2001. Development of floral organ identity: stories from the MADS house. Current Opinion in Plant Biology 4: 75-85.

THEISSEN, G., A. BECKER, A. DI ROSA, A. KANNO, J. T. KIM, T. MUNSTER, K. U. WINTER, AND H. SAEDLER. 2000. A short history of MADS-box genes in plants. Plant Molecular Biology 42: 115-149.

THEISSEN, G., J. T. KIM, AND H. SAEDLER. 1996. Classification and phylogeny of the MADS-box multigene family suggest defined roles of MADS-box gene subfamilies in the morphological evolution of eukaryotes. Journal of Molecular Evolution 43: 484-516.

THEISSEN, G. AND H. SAEDLER. 2001. Plant biology. Floral quartets. Nature 409: 469-471.

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 25: 4876-4882.

TROBNER, W., L. RAMIREZ, P. MOTTE, I. HUE, P. HUIJSER, W. E. LONNIG, H. SAEDLER, H. SOMMER, AND Z. SCHWARZ-SOMMER. 1992. *GLOBOSA*: a homeotic gene which interacts with *DEFICIENS* in the control of *Antirrhinum* floral organogenesis. EMBO Journal 11: 4693-4704.

VAN DE, P. Y., J. S. TAYLOR, I. BRAASCH, AND A. MEYER. 2001. The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. Journal of Molecular Evolution 53: 436-446.

XIANG, J. Q., 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.

XIANG, Q. Y., S. R. MANCHESTER, D. T. THOMAS, W. ZHANG, AND C. FAN. 2005. Phylogeny, biogeography, and molecular dating of cornelian cherries (*Cornus*, Cornaceae): tracking Tertiary plant migration. Evolution 59: 1685-1700.

XIANG, Q. Y. J., D. T. THOMAS, W. H. ZHANG, S. R. MANCHESTER, AND Z. MURRELL. 2006. Species level phylogeny of the genus *Cornus* (Cornaceae) based on molecular and morphological evidence - implications for taxonomy and Tertiary intercontinental migration. Taxon 55: 9-30.

YANG, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Computer Applications in the Biosciences 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. Source of plant materials used in this study and information of clones analyzed. Specimens used in this study including taxa represented, reference identifiers, vouchers. Voucher specimens were collected by Xiang and deposited in NCSC. *: numbers of clones sequenced/numbers of clones analyzed. ‡: sample 02-142 for *PISTILLATA* analysis; sample 02-143 for *APETALA3* analysis.

Subgroups†	Species	Voucher and collection locality	Abbreviation in analyses	*PISTILLATA**	*APETALA3**
BW	*C. alba* L.	02-269, Beijing China	Calba02-269	5/10	11/40
	C. alsophila W.W.Sm.	02-142 (02-143)‡, Yunnan, China	Calso02-142 (Calso02-143)	7/10	5/40
	C. alsophila W.W.Sm.	02-206, Yunnan, China	Calso02-206	7/40	-
	C. alternifolia L.	04-06, West Virginia, USA	Calte04-06	6/40	-
	C. controversa Hemsl.	02-94, Sichuan, China	Ccont02-94	9/49	6/40
	C. oblonga Wall.	02-223, Yunnan, China	Coblo02-223	11/63	8/40
	C. walteri Wangerin	02-267, Beijing, China	Cwalt02-267	8/12	8/40
	C. walteri Wangerin	02-161, Yunnan, China	Cwalt02-161	6/39	-
CC	*C. chinensis* Wangerin	02-83, Sichuan, China	Cchin02-83	6/15	5/40
	C. mas L.	02-282, JC Raulston Arboretum	Cmas02-282	11/39	7/40
	C. officinalis Siebold & Zucc.	918-85A, Arnold Arboretum	Coffi918-85A	10/48	4/40

Table 1. continued.

Subgroups†	Species	Voucher and collection locality	Abbreviation in analyses	*PISTILLATA**	*APETALA3**
BB	*C. angustata* (Chun)T.R.Dudley	02-46, Guangxi, China	Cangu02-46	9/39	9/40
	C. disciflora Sesse & Moc. ex DC.	02-08, Heredia, Costa Rico	Cdisc02-08	11/49	8/40
	C. florida L.	01-1, North Carolina, USA	Cflor01-1	14/62	-
	C. florida L.	01-148, North Carolina, USA	Cflor01-148	11/39	7/40
	C. florida L.	02-16, Javier Clauijero, Costa Rico	Cflor02-16	10/14	-
	C. kousa Buerger ex Miq.	02-268, Beijing, China	Ckous02-268	11/47	-
	C. kousa Buerger ex Miq.	04-C45, Sichuan, China	Ckous04-C45	13/39	6/40
DW	*C. canadensis* L.	Pop6-1, British Columbia, Canada	Ccana6-1	8/41	-
	C. canadensis L.	04-01, West Virginia, USA	Ccana04-01	13/40	10/40
	C. suecica L.	Pop27-2, Alaska, USA	Csuec27-2	9/13	10/40
	C. unalaschkensis Ledeb.	Pop1-1, Washington, USA	Cunal1-1	10/40	-
	C. unalaschkensis Ledeb.	Pop4-1, British Columbia, Canada	Cunal4-1	-	11/40
Outgroups	*Alangium sp.*	02-72, Guangxi, China	Alan02-72	10/22	1/-[d]
	Davidia involucrate Baill.	02-87, Sichuan, China	Davi02-87	10/23	9/40

†BW = blue- or white- fruited group; CC = cornelian cherry; BB = big-bracted group; DW = dwarf dogwoods. d: *Alangium* only has one sequence type of *AP3*-like gene and has clean sequence from direct PCR and sequencing.

Table 2. Comparison of numbers of parsimonious informative sites (# of PIS), percentage of parsimonious informative sites (PPIS), numbers of variable sites (# of VS), and percentage of varible sites (PVS) between floral homeotic paralog genes *PISTILLATA* (*PI*) and *APETALA3* (*AP3*) in *Cornus*.

	Taxa of analysis	*CorPI*		*CorAP3*		*CorPI*		*CorAP3*	
		# of PIS	PPIS (%)	# of PIS	PPIS(%)	# of VS	PVS (%)	# of VS	PVS(%)
Coding	Cornaceae† + outgroups‡	133	65.20	126	58.33	176	86.27	169	78.24
	Cornaceae	72	35.20	91	42.13	109	53.43	124	57.41
	Cornus	62	30.39	48	22.22	103	50.49	81	37.50
Coding and Noncoding	Cornaceae (excluding *Hydrangea*)	-	-	405	55.71	-	-	469	64.51
	Cornus	771	42.02	269	37.00	1017	55.42	338	46.49

†Representatives of Cornaceae included in this study are *Cornus*, *Alangium*, *Davidia* and *Hydrangea*. ‡Outgroups are representatives of core eudicots, lower eudicots and basal angiosperms. Notes, lengths of sequence alignment for coding *CorPI* and entire *CorPI* are 204 bp and 1835 bp, respectively; lengths of sequence alignment for coding *CorAP3* and entire *CorAP3* are 216 bp and 727 bp, respectively. For entire region of *CorPI* gene, only copy *CorPI*-A retained in all *Cornus* species include in this analysis. Since intron

regions are highly variable between *Cornus* and its close relatives *Davidia* and *Alangium,* values of *CorPI* for the category Cornaceae can not be estimated.

Table 3. Numbers of differential sequence types detected in *PISTILLATA*-like and *APETALA3*-like genes in *Cornus*.

	samples	*PISTILLATA*	samples	APETALA-3
BW	Calba02-269	4	Calba02-269	6
	Calso02-142	6	Calso02-143	4
	Calso02-206	6	-	-
	Calte04-06	3	-	-
	Ccont02-94	6	Ccont02-94	3
	Coblo02-223	7	Coblo02-223	6
	Cwalt02-267	6	Cwalt02-267	4
	Cwalt02-161	4	-	-
CC	Cchin02-83	4	Cchin02-83	3
	Cmas02-282	8	Cmas02-282	5
	Coffi918-85A	4	Coffi918-85A	3
BB	Cangu02-46	6	Cangu02-46	1
	Cdisc02-08	4	Cdisc02-08	3
	Cflor01-1	5	-	-
	Cflor01-148	7	Cflor01-148	2
	Cflor02-16	5	-	-
	Ckous02-268	5	-	-
	Ckous04-C45	11	Ckous04-C45	5
DW	Ccana6-1	2	-	-
	Ccana04-01	9	Ccana04-01	7
	Csuec27-2	6	Csuec27-2	6
	Cunal1-1	8	Cunal4-1	3
Outgroups	Alan02-72	4	Alan02-72	1
	Davi02-87	6	Davi02-87	5

For most of the samples, about 40 clones are screened (see Table 1). Sequence differences are counted based on exon analyses. Some clones with identical exon sequences still have

diverged intron regions, e.g., in *C. canadensis* two sequence types have ~100 bp indel difference in intron1.

Table 4. Comparisons of dN/dS ratios of *CorPI* and *CorAP3* between different methods, among different morphological lineages, and between different gene copies.

	CorPI-A		*CorPI*-B		*CorAP3*	
	MEGA	PAML	MEGA†	PAML	MEGA	PAML
Cornus	0.512±0.421	0.496	0.216±0.126	0.351	0.333±0.160	0.256
BB	0.846±0.773	0.496	-	-	0.266±0.211	0.144
DW	1.009±0.501	0.700	-‡	0.251	0.323±0.167	0.309
CC	0.574±0.398	0.344	0.206±0.357	0.374	0.335±0.310	0.239
BW	0.665±0.477	0.519	-	-	0.415±0.316	0.338*

Mean values and standard deviations of dN/dS ratios are estimated based on pairwise comparisons using modified Nei-Gojobori (Jukes-Cantor) model implemented in MEGA 3.1. dN/dS ratios are also estimated using codon-based substitution models implemented in PAML 3.15. BB = big-bracted dogwoods; DW = dwarf dogwoods; CC = cornelian cherries; BW = blue- or white- fruited dogwoods. †: BB and BW subgroups lost the *CorPI*-B copy. ‡: the dS values of DW are all zero and dN values are 0~0.018. *: excluding *C. oblonga*.

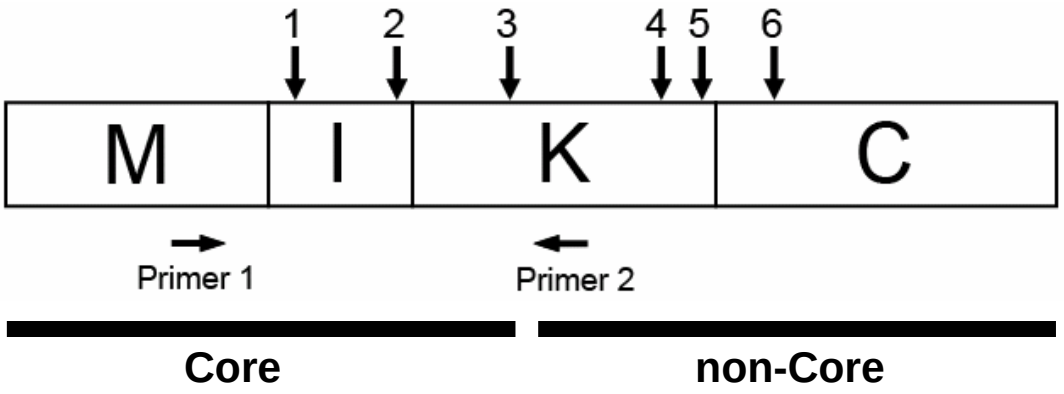


Figure 1. Schematic diagram of MIKC structure of floral homeotic B-class genes. Region between primer 1 and primer 2 is analyzed. Positions of six introns are indicated. Core and non-core regions are indicated based on functional studies.

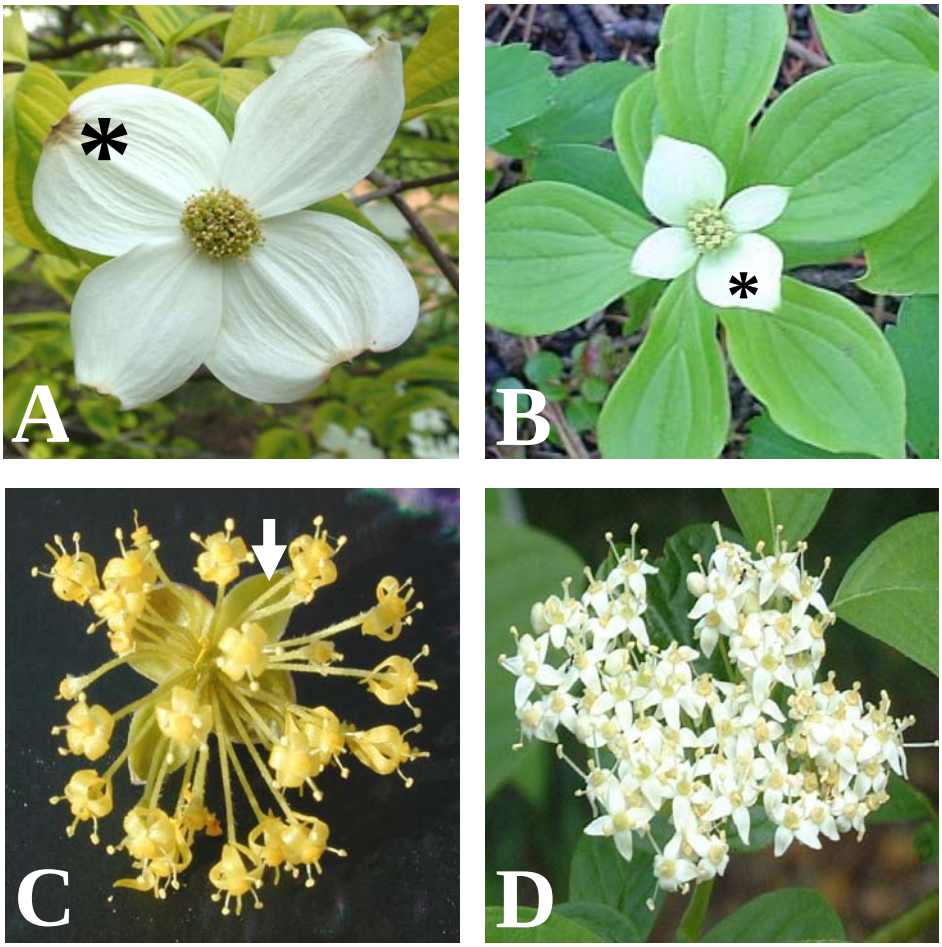


Figure 2. A, *Cornus florida* (Big-bracted dogwoods, BB). B, *Cornus canadensis* (Dwarf dogwoods, DW). C, *Cornus mas* (Cornelian cherries, CC). D, *Cornus* spp. (Blue- or white-fruited dogwoods, BW). Asterisks indicate bracts, which are petaloid in A and B; arrow in dicates bracts, which are modified but not petaloid in C; bracts are leaf-like and early deciduous in D.

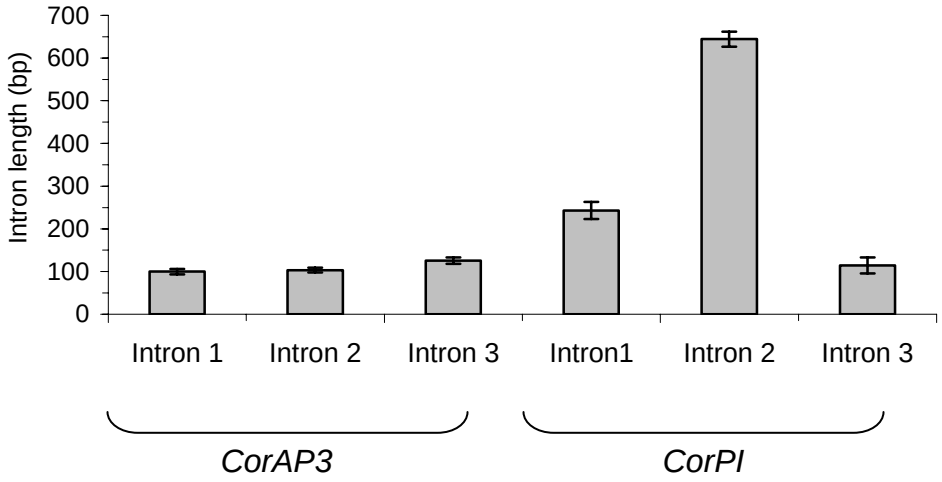


Figure 3. Length variation of intron regions in *CorPI* and *CorAP3* in *Cornus*.

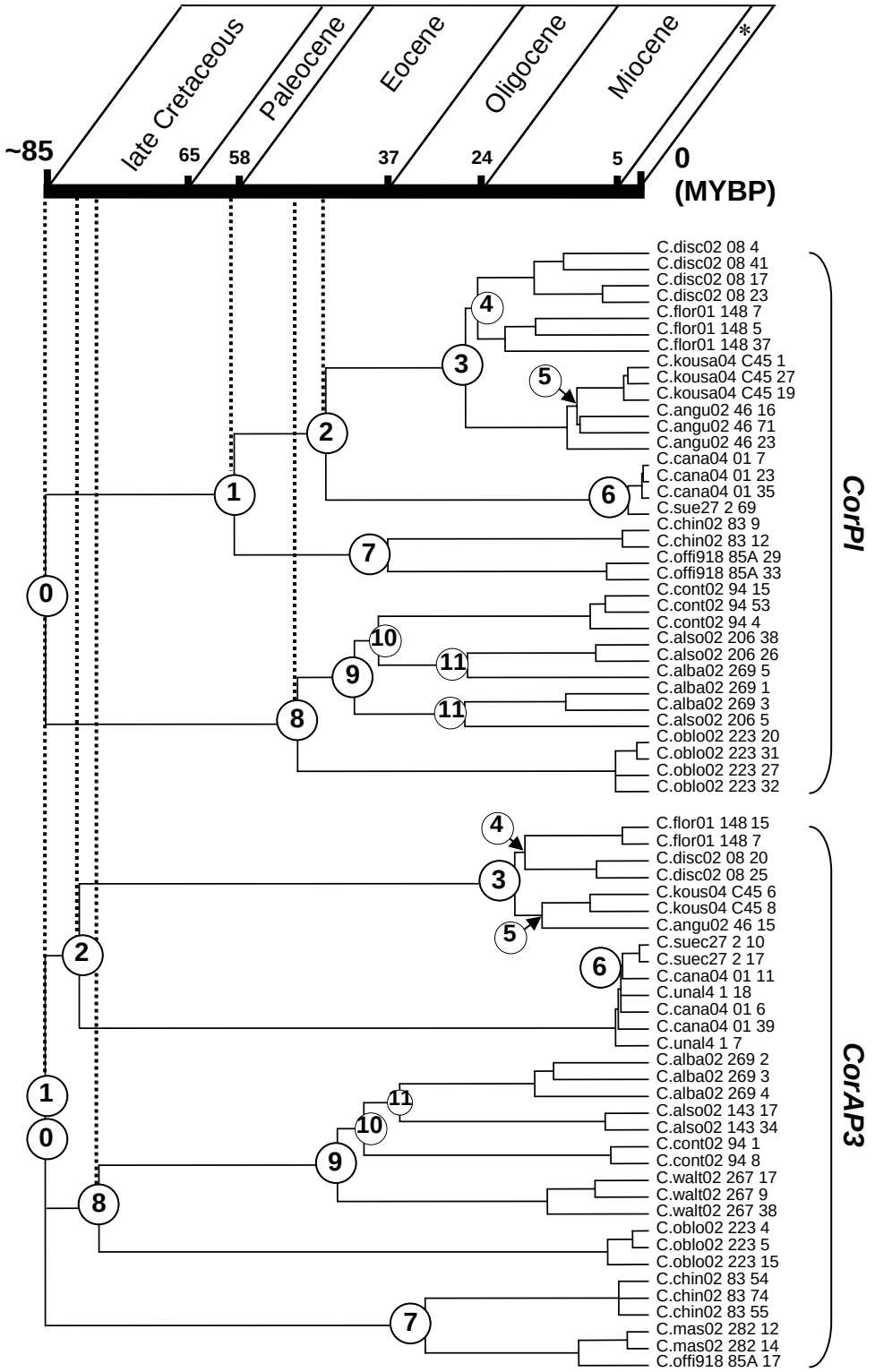

late Cretaceous
Paleocene
Eocene
Oligocene
Miocene
*
~85
65
58
37
24
5
0
(MYBP)
C.disc02 08 4
C.disc02 08 41
C.disc02 08 17
C.disc02 08 23
C.flor01 148 7
C.flor01 148 5
C.flor01 148 37
C.kousa04 C45 1
C.kousa04 C45 27
C.kousa04 C45 19
C.angu02 46 16
C.angu02 46 71
C.angu02 46 23
C.cana04 01 7
C.cana04 01 23
C.cana04 01 35
C.sue27 2 69
C.chin02 83 9
C.chin02 83 12
C.offi918 85A 29
C.offi918 85A 33
C.cont02 94 15
C.cont02 94 53
C.cont02 94 4
C.also02 206 38
C.also02 206 26
C.alba02 269 5
C.alba02 269 1
C.alba02 269 3
C.also02 206 5
C.oblo02 223 20
C.oblo02 223 31
C.oblo02 223 27
C.oblo02 223 32
CorPI
C.flor01 148 15
C.flor01 148 7
C.disc02 08 20
C.disc02 08 25
C.kous04 C45 6
C.kous04 C45 8
C.angu02 46 15
C.suec27 2 10
C.suec27 2 17
C.cana04 01 11
C.unal4 1 18
C.cana04 01 6
C.cana04 01 39
C.unal4 1 7
C.alba02 269 2
C.alba02 269 3
C.alba02 269 4
C.also02 143 17
C.also02 143 34
C.cont02 94 1
C.cont02 94 8
C.walt02 267 17
C.walt02 267 9
C.walt02 267 38
C.oblo02 223 4
C.oblo02 223 5
C.oblo02 223 15
C.chin02 83 54
C.chin02 83 74
C.chin02 83 55
C.mas02 282 12
C.mas02 282 14
C.offi918 85A 17
CorAP3

Figure 4. Comparison of divergence time estimations based on *CorPI-A* and *CorAP3*. Numbers indicated in nodes correspond to speciation events. node 0: origin of *Cornus*; node 1: split of CC from the rest; node 2: split of DW and BB; node 3: split of BB-American and BB-Asian; node 4: split of *C. florida*-North America and *C. disciflora*-South America; node 5: split of *C. kousa* and *C. angustata*; node 6: split *C. canadensis* and *C. suecica*; node 7: split of *C. chinensis* and *C. officinalis*; node 8: split of *C. oblonga* from BW; node 9: divergence of BW lineage excluding *C. oblonga*; node 10: *C. controversa* branch out; node 11: split of *C. alsophila* and *C. alba*. Nodes 1, 2 and 8 are specifically emphasized due to significant differences of estimations based on *CorPI* and *CorAP3* genes. *: < 5 million year ago, Pliocene and Quaternary.

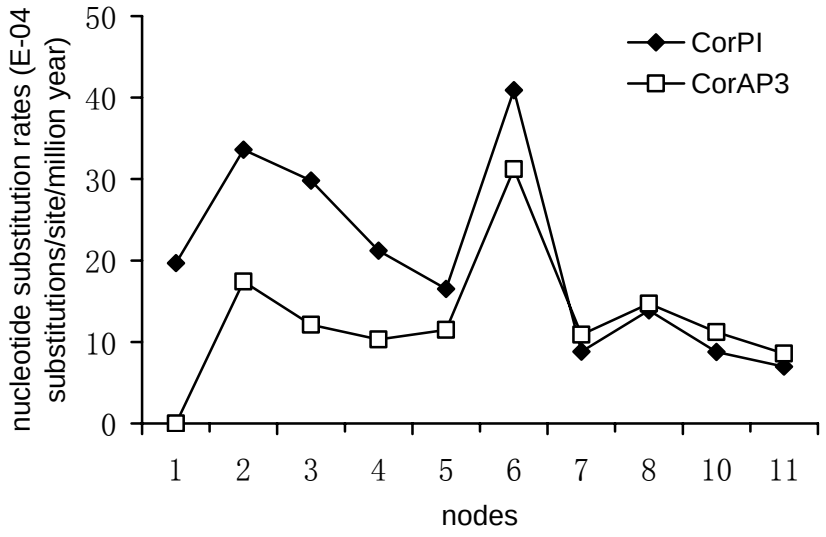


Figure 5. Comparison of nucleotide substitution rates along corresponding branches based on entire *CorPI-A* and *CorAP3* genes. Node numbers correspond to those in Fig 3.