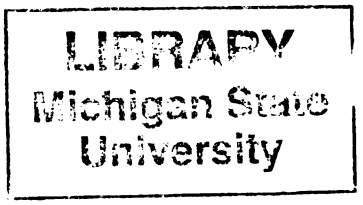


This is to certify that the

thesis entitled

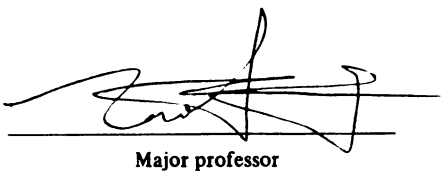
EVOLUTION AND PHYLOGENETIC UTILITY OF LOW-COPY NUCLEAR GENES: EXAMPLES FROM CONIFERS AND PEONIES

presented by

David C. Tank

has been accepted towards fulfillment
of the requirements for

M.S. degree in Botany & Plant Pathology



Major professor

Date May, 9, 2000

EVOLUTION AND PHYLOGENETIC UTILITY OF LOW-COPY NUCLEAR GENES: EXAMPLES FROM CONIFERS AND PEONIES

By

David C. Tank

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE

Department of Botany and Plant Pathology

2000

ABSTRACT

EVOLUTION AND PHYLOGENETIC UTILITY OF LOW-COPY NUCLEAR GENES: EXAMPLES FROM CONIFERS AND PEONIES

By

David C. Tank

Low-copy nuclear genes have the potential to provide multiple, independent gene phylogenies that can be used to reconstruct species phylogenies, and may be more appropriate for resolving low-level phylogenetic relationships, such as those among closely related species, than common molecular phylogenetic markers. The goals of this study were to 1) investigate the molecular evolution of low-copy nuclear genes in a phylogenetic context, and 2) investigate the phylogenetic utility of low-copy nuclear genes through comparison to previous phylogenetic hypotheses. To obtain these goals, example low-copy nuclear gene markers were examined in the conifer families Pinaceae and Taxodiaceae, and the angiosperm genus *Paeonia* (Paeoniaceae).

The nuclear genomes of most conifers are large and organized in complex gene families. The gene encoding cinnamyl alcohol dehydrogenase (CAD) is a key enzyme in the lignin biosynthetic pathway. Three main types of the CAD gene were identified by neighbor-joining analysis. Type I CAD consists of sequences isolated from Pinaceae species only, and were determined to be mostly orthologous and evolving at a rate representative of the rate of nuclear gene divergence in Pinaceae. In both type II and III CAD neither Pinaceae nor Taxodiaceae sequences are monophyletic, and sequence divergence within Taxodiaceae, and between the two families, is extremely variable. Based on comparisons to other genes, the type II and III CAD divergences were

determined to be as much as 214-times and 256-times lower than expected, within Taxodiaceae, and between Pinaceae and Taxodiaceae, respectively. Two hypotheses are proposed to explain the results: 1) extensive paralogy within and between type II and III CAD, combined with an extremely low divergence rate at some of the paralogous loci, and 2) lateral gene transfer both between genera of Taxodiaceae, and between the two conifer families. If the first hypothesis is invoked, the rate of divergence between some CAD genes would have to be as low as 9.6×10^{-12} substitutions/site/year. This is > 600× less than previous estimates of synonymous sequence divergence in plant nuclear genes. As there is no known evolutionary mechanism that can explain the maintenance of such a strikingly low sequence divergence rate, we feel that it is more likely the observed divergence patterns are the result of lateral gene transfer between species.

The nuclear encoded chloroplast-expressed glycerol-3-phosphate acyltransferase gene (GPAT) has been found to be single copy in a number of angiosperm families. In this study we investigated 1) the molecular evolution of the GPAT gene in *Paeonia* through comparison to previous phylogenetic hypotheses, and 2) the phylogenetic utility of the GPAT gene in *Paeonia*. An approximately 2.3-2.6 kb fragment of the GPAT gene was amplified, cloned, and sequenced from 13 *Paeonia* species. Parsimony analysis resolved a highly supported GPAT gene phylogeny that differed from previous phylogenetic hypotheses in two areas. When the topology of the GPAT phylogeny was evaluated with the Templeton test, one discordance was determined to be significantly incongruent. Two distinct genomic clones of *P. anomala* containing the GPAT gene have been characterized and suggest that the gene underwent an ancient duplication event followed by the formation of a pseudogene in one copy. BLAST sequence similarity

analysis suggests that the GPAT pseudogene may contain a large retrotransposon-like insertion that may have been the gene silencing mechanism of this locus. These results suggest that, unlike the GPAT gene history in other angiosperms, in Peonies the GPAT gene may have undergone duplication and deletion. While the GPAT gene is useful for phylogeny reconstruction at a 'local' level in *Paeonia*, it may present paralogous relationships when investigating the relationships within the genus as a whole.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Tao Sang, for his support and guidance throughout my academic career at Michigan State University. Tao has been a true mentor, and without his support and enthusiasm this research could not have been accomplished. I must also thank Tao for taking me under his wing when I was an undergraduate. Without his guidance at that time my academic future would not be what it is today. In addition, many thanks to the members of my graduate committee, Drs. Alan Prather and Gerry Adams, for their guidance, support, and criticism throughout. I would also like to thank Dr. Jeffery White for his natural ability to teach. Without Jeff I would not have found my niche.

While in Tao's lab I have had the opportunity to work with a number of outstanding researchers who helped create an intellectually stimulating and enjoyable working environment. I thank Drs. Xiao-Quan Wang, Diane Ferguson, and Song Ge for their helpful discussions, assistance in the lab, and good cheer. I especially thank Xiao-Quan for his incredible knowledge of molecular techniques and his willingness to convey that knowledge to me. Without Xiao-Quan much of the conifer research could not have been completed.

Finally, I owe everything to my family and friends. I would like to thank specifically my parents for their love and support in everything I do, and my best friend Kara for sticking with me all these years, and putting up with me through the stressful times of deadlines.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

One of the primary goals of molecular phylogenetic studies is the reconstruction of species phylogenies from separate and combined analyses of individual gene phylogenies. Commonly used molecular markers in plant phylogenetic studies include genic and intergenic regions of both chloroplast DNA (cpDNA) and nuclear ribosomal DNA (nrDNA), both of which exist in high copy numbers in plant cells. Chloroplast DNA lacks intracellular variation or recombination, resembling that of a single-copy gene. Likewise, due to concerted evolution of gene members, sequences of nrDNA usually lack polymorphism in an individual. Therefore, the PCR pool of either cpDNA or nrDNA is usually homogeneous, and PCR products can be sequenced directly.

However, because sequence divergence rates in both cpDNA and nrDNA are generally low, molecular markers from these regions are often not appropriate for investigating low-level phylogenetic relationships in plants, such as those among closely related species. Furthermore, neither cpDNA nor nrDNA is useful for reconstructing hybrid speciation, as cpDNA is generally maternally inherited, and nrDNA is homogenized through concerted evolution following hybridization. In addition, gene phylogenies from cpDNA and nrDNA sequence data are often conflicting (e.g., Soltis and Kuzoff 1995; Maon-Gamer and Kellogg 1996; Sang, Crawford and Stuessy 1997), and limited numbers of independent gene phylogenies will impede attempts to reconstruct accurate species phylogenies. Therefore, it is necessary to obtain additional independent gene phylogenies to reconstruct stronger hypotheses of the one underlying phylogeny - the species phylogeny.

Low-copy nuclear genes have the potential to provide an abundance of independent gene phylogenies. Aside from the sheer number of potential independent markers, low-copy nuclear genes are biparentally inherited, and generally diverge at a higher rate, most notably in intron regions, than cpDNA or nrDNA. Therefore, low-copy nuclear genes are especially useful for reconstructing low-level taxonomic relationships in which cpDNA and nrDNA nucleotide sequences are too conserved to resolve.

The use of low-copy nuclear genes in molecular phylogenetic studies of plants is increasing (e.g., Gottlieb and Ford 1996; Doyle, Kanazin, and Shoemaker 1996; Sang, Donoghue, and Zhang 1997; Mason-Gamer, Weil, and Kellogg 1998; Small et al. 1998; Emshwiller and Doyle 1999; Matthews and Donoghue 1999; Wang, Tank, and Sang 2000). However, in comparison to more commonly used molecular markers in plant systematic studies (i.e., cpDNA and nrDNA genes and spacers), the phylogenetic utility of low-copy nuclear genes is still largely understudied. This is due primarily to difficulties in determining orthology from paralogy among members of a gene family, and the increased lab-work necessary for cloning. The selection of genes that exist in relatively small gene families, and that are less dynamic in duplication and deletion can aid in overcoming these difficulties.

The overall objectives of the following studies were to 1) investigate the molecular evolution of low-copy nuclear genes in a phylogenetic context, including the dynamics of duplication and deletion of low-copy nuclear loci, and mechanisms of low-copy nuclear gene evolution causing discordance among gene phylogenies, and 2) investigate the phylogenetic utility of low-copy nuclear genes through comparison to previous phylogenetic hypotheses. To obtain these goals, example low-copy nuclear

gene markers were examined in the conifer families Pinaceae and Taxodiaceae, and the angiosperm genus *Paeonia* (Paeoniaceae).

CHAPTER 1

DIFFERENCES IN COPY NUMBER SUGGEST ATYPICAL EVOLUTION OF THE CINNAMYL ALCOHOL DEHYDROGENASE GENE IN THE CONIFER FAMILIES PINACEAE AND TAXODIACEAE

INTRODUCTION

The nuclear genomes of most conifers are large and organized in complex gene families (Kinlaw and Neale 1997; Murray 1998). Very little research has been done to investigate the dynamics of nuclear gene evolution in conifers, and the question of how such complexity at both the genomic and genic level has arisen in conifers remains open (Kinlaw and Neale 1997).

Cinnamyl alcohol dehydrogenase (CAD) regulates the last step of lignin biosynthesis by catalyzing the reduction of cinnamaldehydes to cinnamyl alcohols. This reduction occurs after the branch points between the lignin biosynthetic pathway and the pathway for phenylpropanoid metabolism for flavenoids and other phenolic compounds (O'Malley, Porter, and Sederoff 1992; MacKay et al. 1997). For this reason, CAD has been considered the 'molecular marker' for lignin biosynthesis (Walter et al. 1988). The CAD gene is present as a single copy in loblolly pine (*Pinus taeda* L.; O'Malley, Porter, and Sederoff 1992; MacKay et al. 1995; MacKay et al. 1997), but exists in at least two copies in Norway spruce (*Picea abies* L.; Schubert et al. 1998), a member of the closely

related genus *Picea*. This suggests that the CAD gene is a good marker for investigating the dynamics of low-copy nuclear gene evolution in conifers.

Single- and low-copy nuclear genes are being used more frequently in phylogenetic analyses of angiosperms (e.g., Gottlieb and Ford 1996; Doyle, Kanazin, and Shoemaker 1996; Sang, Donoghue, and Zhang 1997; Mason-Gamer, Weil, and Kellogg 1998; Small et al. 1998; Emshwiller and Doyle 1999; Matthews and Donoghue 1999). Recently, the low-copy nuclear gene encoding 4 courmarate : coenzyme A ligase (4CL), an enzyme also found in the lignin biosynthetic pathway, was used to infer the phylogeny of Pinaceae (Wang, Tank, and Sang 2000). 4CL provided a wealth of phylogenetically informative characters that made it possible to reconstruct a well-resolved and supported intergeneric phylogeny.

Pinaceae, the largest extant family of gymnosperms, is comprised of 11 genera and more than 200 species (Farjon 1998). Pinaceae is both ecologically and economically important, as many members of the family constitute the major forest elements of the northern temperate region. Phylogenetic analyses of chloroplast DNA indicate that the sister family of Pinaceae is Taxodiaceae (Chase et al. 1993; Brunsfeld et al. 1994; Tsumura et al. 1995). Taxodiaceae consists of 10 genera and only ~14 species. Because of its great diversity and wide geographic distribution in the fossil record (Miller 1977), and the present abundance of endemic and monotypic genera, Taxodiaceae is often considered a relictual family (Brunsfeld et al. 1994). Both Pinaceae and Taxodiaceae are complimented by an extensive fossil record that supports the divergence of the two families at least 200 million years ago (Florin 1963).

The primary objective of this study was to investigate the molecular evolution of the CAD gene in the conifer families Pinaceae and Taxodiaceae. To investigate the evolutionary dynamics of the CAD gene family, a neighbor-joining (NJ) analysis was conducted with partial sequences of the gene obtained by polymerase chain reaction (PCR). In addition, CAD sequence divergence within and between the two conifer families was estimated and compared.

MATERIALS AND METHODS

All 11 recognized genera of Pinaceae were sampled, including *Abies* (fir), *Cathaya*, *Cedrus* (cedar), *Keteleeria*, *Larix* (larch), *Nothotsuga*, *Picea* (spruce), *Pinus* (pine), *Pseudolarix* (golden larch), *Pseudotsuga* (Douglas-fir), and *Tsuga* (hemlock). Sampling of Taxodiaceae was limited to five of the 10 recognized genera, including *Cryptomeria*, *Metasequoia* (dawn redwood), *Sequoia* (coast redwood), *Sequoiadendron* (giant sequoia), and *Taxodium* (bald cypress). Sampling localities are given in Table 1-1, and voucher specimens have been deposited in the herbaria of the Institute of Botany, Beijing and Michigan State University. Total DNA was isolated from fresh leaves using the CTAB method (Doyle and Doyle 1987) and purified with a Wizard DNA Clean-up System (Promega).

PCR AND SEQUENCING

The CAD gene was amplified through the following PCR cycles: (1) 70°C, 4 min; (2-4) 94°C, 1 min; 48-55°C, 30 sec; 72°C, 2 min; (5-7) 94°C, 20 sec; 48-55°C, 30 sec; 72°C, 2 min (repeat 5-7 29 times); (8) 72°C, 10 min. The forward primers CAD40F (5'-CAGCTCGGGACTCCAGTGG) and CADF2 (5'-CCTTACACTTACAATCTCAG), located on exon 1, and CADF3 (5'-GTCAGGGTCATTTACTGCGG), located on exon 2,

Table 1-1. Collection locality of species of Pinaceae and Taxodiaceae sampled for DNA sequencing

Species	Collection locality
Abies beshanzuensis Wu	Longquan, Zhejiang, China
Abies firma Sieb. et Zucc.	Botanic Garden, Institute of Botany, Beijing
Abies holophylla Maxim.	Botanic Garden, Institute of Botany, Beijing
Cathaya argyrophylla Chun et Kuang	Huaping, Guangxi, China
Cedrus atlantica Manetti	Michigan State University, East Lansing
Keteleeria evelyniana Mast.	Botanic Garden, Institute of Botany, Kunming
Larix gmelini (Rupr.) Rupr.	Botanic Garden, Institute of Botany, Beijing
Nothotsuga longibracteata Hu ex Page	Xinning, Hunan, China
Picea smithiana (Wall.) Boiss.	Botanic Garden, Institute of Botany, Beijing
Pinus armandi Franch.	Botanic Garden, Institute of Botany, Beijing
Pinus banksiana Lamb.	Botanic Garden, Institute of Botany, Beijing
Pseudolarix amabilis (Nelson) Rehd.	Botanic Garden, Institute of Botany, Kunming
Pseudotsuga menziesii (Mirbel) Franco	Botanic Garden, Institute of Botany, Beijing
Pseudotsuga sinensis Dode	Botanic Garden, Institute of Botany, Kunming
Tsuga canadensis Carr.	Michigan State University, East Lansing
Tsuga mertensiana (Bong.) Rydb.	Mt. Hood, OR
Cryptomeria japonica D. Don	Michigan State University, East Lansing
Metasequoia glyptostroboides Hu et Chang	Botanic Garden, Institute of Botany, Beijing (1)
" "	Michigan State University, East Lansing

Table 1-1 (cont'd)

Sequoia sempervirens (D. Don) Endl.	Rancho Santa Ana Botanic Garden, CA
Sequoiadendron giganteum (Lindl.) Buchholz	Rancho Santa Ana Botanic Garden, CA
Taxodium distichum (L.)	Michigan State University, East Lansing

and the reverse primers CAD1.5R (5'-AACGGCTCTGGAACAACGCC), CADR2 (5'-GGGCAACTGGAATGGTGTC), and CADR4 (5'-CTAGGCTCTCTGCTGCTTCC) located on exon 5, were used to amplify a portion of the CAD gene from all Pinaceae and Taxodiaceae species (Figure 1-1). These primers were designed in the most conservative regions found between both conifer and angiosperm CAD sequences available in genbank. To amplify CAD from all accessions, it was necessary to design multiple sets of CAD primers. Combinations of the three forward and reverse primers were tried until amplification was successful.

Amplified PCR products were cloned with a TA cloning kit (Invitrogen). For each species, 10 to 30 clones were screened by examining restriction-site or sequence (from one primer) variation (Sang, Donoghue, and Zhang 1997; Wang, Tank, and Sang 2000). Distinct clones were fully sequenced and included in the phylogenetic analyses. Sequencing was done on an ABI 373 automated DNA sequencer using either the Dye Terminator Cycle Sequencing reaction kit (PE Applied Biosystems) or the DYEnamic ET Terminator Cycle Sequencing reaction kit (Amersham Pharmacia Biotech).

Preliminary phylogenetic analysis indicated three main types of the CAD gene in Pinaceae and Taxodiaceae. To assure that all types of CAD present in each accession were isolated, further PCR screening was conducted using newly designed type specific CAD primers (Table 1-2, Figure 1-1). Resulting PCR products from type specific amplifications were cloned, and 10-30 clones were isolated and screened from both Pinaceae and Taxodiaceae species. Distinct clones were fully sequenced and used in the phylogenetic analyses. Upon submission for publication all CAD sequences will be deposited in GenBank. Additional CAD sequences obtained from GenBank for this

Table 1-2. Type specific CAD primers used for PCR screening and isolation of type I, II and III CAD

Primer	Location	Sequence	Specificity
CADSPRp	exon 5	5'-TCCATTTGTCTTTAGAAGGGC	type I
CADNOF	exon 2	5'-CTGCCACTCTGACTTATCGG	type IIA
CADSPFa	exon 1	5'-ACACTCTCAGGTACATCTATCC	type IIB
CADSPRa	exon 5	5'-TCCATTTGTCTTTARAAGGGA	type IIB
CADSLR1	exon 4	5'-CTGTCATACCGAAATGCTTCAA	type III
CADSLR2	intron 4	5'-CTGCAAGACAACGGATTCACTT	type III

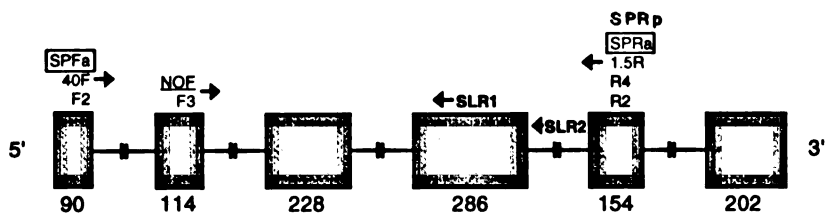


Figure 1-1. Structure of the CAD gene in Pinaceae and Taxodiaceae. Boxes represent exon regions with the corresponding length in base pairs underneath each exon. Intron regions are characterized as broken lines between exons, as intron length is variable. All CAD primers used for PCR amplification are labeled above the exon in which they were designed: plain text, general primers; shadowed text, type I specific; underlined text, type IIA specific; boxed text, type IIB specific; bold text, type III specific.

study include: *Picea abies* 2 (AJ001924), *Picea abies* 7 (AJ001925), *Picea abies* 8 (AJ001926; Schubert et al. 1998), and *Pinus radiata* (AF060491; Moyle, Wagner, and Walter 1998).

Data Analyses

Sequence alignments were made with ClustalW (Thompson, Higgins, and Gibson 1994) and refined manually. Regions in the CAD introns that could not be aligned unambiguously were excluded from analyses. Neighbor-joining (NJ) analysis, as implemented in PAUP* 4.0 (Swofford 1998), was used to infer phylogenetic relationships based on nucleotide substitutions in aligned sequences. NJ was selected for phylogenetic analyses because it is less sensitive to rate heterogeneity, and therefore is more consistent than parsimony in cases of extreme rate heterogeneity as observed here (Huelsenbeck and Hillis 1993). Genetic distances for the NJ analyses were estimated via maximum-likelihood using the model of sequence evolution that best fit the data set by the hierarchical likelihood ratio test, as determined with the program Modeltest 2.1 (Posada and Crandall 1998). The resulting NJ tree was rooted with the monophyletic type III CAD sequence group (Figure 1-2). Support for each node was calculated by the NJ bootstrap method with 1000 replicates of random taxon addition, as implemented in PAUP* 4.0. Sequence divergence was estimated using Jukes-Cantor corrections (Jukes and Cantor 1969) for all nucleotides, as calculated by PAUP* 4.0, and for synonymous and nonsynonymous sites, as calculated by MEGA 1.02 (Kumar, Tamura, and Nei 1993).

Figure 1-2. Neighbor-joining tree of all CAD sequences isolated from Pinaceae and Taxodiaceae species. Distances were calculated via maximum-likelihood using the Tamura-Nei (1993) model of sequence evolution. Substitution rates were assumed to follow a gamma distribution with the shape parameter estimated via maximum-likelihood (.606153). Numbers associated with species names correspond to clone numbers, numbers associated with branches correspond to bootstrap support >50%, branch lengths are proportional to genetic distance as measured by the scale bar. Monophyletic groups have been further categorized as type I, II (A or B), or III as indicated. Clones shown in bold are pseudogenes.

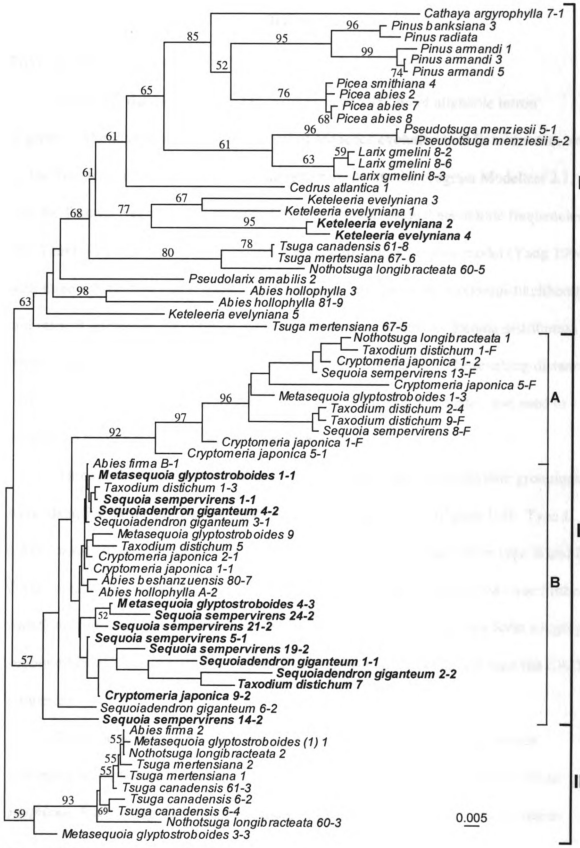

Cathaya argyrophylla 7-1
Pinus banksiana 3
Pinus radiata
Pinus armandi 1
Pinus armandi 3
Pinus armandi 5
Picea smithiana 4
Picea abies 2
Picea abies 7
Picea abies 8
Pseudotsuga menziesii 5-1
Pseudotsuga menziesii 5-2
Larix gmelini 8-2
Larix gmelini 8-6
Larix gmelini 8-3
Cedrus atlantica 1
Keteleeria evelyniana 3
Keteleeria evelyniana 1
Keteleeria evelyniana 2
Keteleeria evelyniana 4
Tsuga canadensis 61-8
Tsuga mertensiana 67- 6
Nothotsuga longibracteata 60-5
Pseudolarix amabilis 2
Abies hollophylla 3
Abies hollophylla 81-9
Keteleeria evelyniana 5
Tsuga mertensiana 67-5
Nothotsuga longibracteata 1
Taxodium distichum 1-F
Cryptomeria japonica 1-2
Sequoia sempervirens 13-F
Cryptomeria japonica 5-F
Metasequoia glyptostroboides 1-3
Taxodium distichum 2-4
Taxodium distichum 9-F
Sequoia sempervirens 8-F
Cryptomeria japonica 1-F
Cryptomeria japonica 5-1
Abies firma B-1
Metasequoia glyptostroboides 1-1
Taxodium distichum 1-3
Sequoia sempervirens 1-1
Sequoiadendron giganteum 4-2
Sequoiadendron giganteum 3-1
Metasequoia glyptostroboides 9
Taxodium distichum 5
Cryptomeria japonica 2-1
Cryptomeria japonica 1-1
Abies beshanzuensis 80-7
Abies hollophylla A-2
Metasequoia glyptostroboides 4-3
Sequoia sempervirens 24-2
Sequoia sempervirens 21-2
Sequoia sempervirens 5-1
Sequoia sempervirens 19-2
Sequoiadendron giganteum 1-1
Sequoiadendron giganteum 2-2
Taxodium distichum 7
Cryptomeria japonica 9-2
Sequoiadendron giganteum 6-2
Sequoia sempervirens 14-2
Abies firma 2
Metasequoia glyptostroboides (1) 1
Nothotsuga longibracteata 2
Tsuga mertensiana 2
Tsuga mertensiana 1
Tsuga canadensis 61-3
Tsuga canadensis 6-2
Tsuga canadensis 6-4
Nothotsuga longibracteata 60-3
Metasequoia glyptostroboides 3-3
0.005
I
A
II
B
III

Figure 1-2

RESULTS

PHYLOGENETIC ANALYSIS

The CAD data set contains 738 bp of exon and 357 bp of alignable intron sequence. The maximum-likelihood model of sequence evolution best fit to the data set by the hierarchical likelihood ratio test, as determined with the program Modeltest 2.1, was the Tamura-Nei model (Tamura and Nei 1993) with unequal nucleotide frequencies. The Tamura-Nei model is a sub-model of the general-time-reversible model (Yang 1994) with three substitution types, nucleotide frequencies estimated via maximum-likelihood, and rates of nucleotide substitution assumed to follow a continuous gamma distribution (shape parameter = 0.60615, estimated via maximum-likelihood). The resulting distance matrix, calculated via maximum-likelihood using the Tamura-Nei model, was used to construct a NJ tree from the CAD data set (Figure 1-2).

Three main types of the CAD gene, recognized by their monophyletic groupings, were identified by the NJ analysis, and are labeled on the NJ tree (Figure 1-2). Type I CAD consists of only sequences isolated from members of Pinaceae, while type II and III CAD contain sequences from both Taxodiaceae and Pinaceae. Type II CAD was further partitioned into type IIA and IIB CAD, in which type IIA CAD sequences form a highly supported (92% bootstrap support) monophyletic group nested within the type IIB CAD sequences.

Pinaceae species in which only type I CAD clones were identified include *Cathaya argyrophylla*, *Cedrus atlantica*, *Keteleeria evelyniana*, *Larix gmelini*, *Picea smithiana*, *Pinus armandi*, *Pinus banksiana*, *Pseudolarix amabilis*, and *Pseudotsuga menziesii*. In addition to type I CAD sequences identified in *Abies holophylla*,

Nothotsuga longibracteata, *Tsuga canadensis*, and *T. mertensiana*, type IIA (*N. longibracteata*), IIB (*A. holophylla*) and type III (*N. longibracteata*, *T. canadensis*, and *T. mertensiana*) CAD clones were also identified from these species. The only Pinaceae species in which a type I CAD gene was not identified were *Abies beshanzuensis* and *A. firma*, in which only type IIB CAD was isolated. Species of Taxodiaceae were found to contain both type IIA and IIB CAD sequences, with the exception of *Sequoiadendron giganteum*, in which only type IIB CAD sequences were identified, and *Metasequoia glyptostroboides*, in which type IIA, IIB, and type III CAD clones were found.

For the most part, when multiple distinct type I CAD sequences were found for a genus, they formed well-supported monophyletic groups on the NJ tree (e.g., *Abies holophylla*, *Larix gmelini*, *Pinus* sp., *Picea*, sp., and *Pseudotsuga menziesii*; Figure 1-2). For both *Keteleeria* and *Tsuga*, one type I CAD sequence is resolved at the base of the type I group, however, the position of these two clones (*Keteleeria evelyniana* 5 and *Tsuga mertensiana* 67-5; Figure 1-2) within type I CAD is not supported by bootstrap values. To determine whether the type I CAD sequences are orthologous among genera of Pinaceae, all type I CAD sequences were subjected to a parsimony analysis with Cedrus as the functional outgroup, as indicated by previous phylogenetic analyses (Wang, Tank, and Sang 2000). A heuristic search with 1000 replicates of random taxon addition resulted in four most parsimonious trees (trees not shown). Unlike the NJ tree (Figure 1-2), on the type I CAD parsimony trees all clones isolated from a genus form well-supported monophyletic clades, including clones isolated from both *Keteleeria* and *Tsuga*. The strict consensus of the parsimony trees is almost identical to the 'species phylogeny' of Pinaceae, inferred previously from a combined parsimony analysis of gene

sequences isolated from each of the three genomes (Wang, Tank, and Sang 2000), except for the position of *Pseudolarix*. Using the Pinaceae 'species phylogeny' as a topological constraint, both the Templeton test ($p = 0.3750$; Templeton 1983) and the Kishino-Hasegawa test ($p = 0.5226$; Kishino and Hasegawa 1989) indicate that the incongruence is not significant.

Analysis of Sequence Divergence

To evaluate the amount of sequence divergence between orthologous type I CAD sequences among genera of Pinaceae, mean synonymous and nonsynonymous sequence divergence was estimated for a type I CAD data set reduced to 13 sequences, with at least one clone representing each of the 11 genera. In genera where multiple type I CAD sequences were isolated, one clone was randomly chosen to represent each, with the exception of *Pinus* and *Tsuga*, in which two species were selected to represent each genus. These divergences were compared to the mean synonymous and nonsynonymous sequence divergence of the 4CL gene for an almost identical set of species used previously to construct the combined three-genome phylogeny of Pinaceae (Wang, Tank, and Sang 2000). Those clones used in the analysis include: *Abies holophylla* 81-9, *Cathaya argyrophylla* 7-1, *Cedrus atlantica* 1, *Keteleeria evelyniana* 1, *Larix gmelini* 8-3, *Nothotsuga longibracteata* 60-5, *Picea smithiana* 4, *Pinus armandi* 3, *P. banksiana* 3, *Pseudolarix amabilis* 2, *Pseudotsuga menziesii* 5-1, *Tsuga canadensis* 61-8, and *T. mertensiana* 67-6. The mean synonymous and nonsynonymous sequence divergence estimates for the 4CL data set were 0.3224 ± 0.0265 and 0.0540 ± 0.0054, respectively, and the mean synonymous and nonsynonymous divergences estimated for the reduced CAD type I data set were 0.2845 ± 0.0244 and 0.0499 ± 0.0051, respectively. The mean

synonymous and nonsynonymous sequence divergence estimates were not significantly different for the two data sets ($p > 0.05$).

In sharp contrast to the type I CAD sequences, the evolutionary dynamics of type II and III CAD for both the Pinaceae and Taxodiaceae are quite different. There are a few striking points to be mentioned. First, in both type II and type III CAD, neither Pinaceae nor Taxodiaceae are resolved as a monophyletic group. The type II CAD group is comprised of mostly sequences from Taxodiaceae species, however, one *Nothotsuga longibracteata* CAD clone is nested within type IIA CAD, and one clone from each of the three *Abies* species included in the analysis is nested within the type IIB group (Figure 1-2). The type III CAD group consists of 10 CAD clones, eight isolated from *Abies*, *Nothotsuga*, and *Tsuga*, and two clones identified in *Metasequoia* (one from each accession).

Second, the amount of type II and III CAD sequence divergence within Taxodiaceae, and between the two families, is extremely variable. To illustrate the variability in sequence divergence *Metasequoia glyptostroboides* and *Cryptomeria japonica* were compared within Taxodiaceae, and between the two conifer families *Metasequoia glyptostroboides* and *Abies* were chosen to represent Taxodiaceae and Pinaceae, respectively. Within Taxodiaceae, the divergence between all pairwise comparisons of type II and III CAD sequences from *Metasequoia glyptostroboides* and *Cryptomeria japonica* was calculated (Figure 1-3). These divergence values differed by as much as 34-fold, ranging from 0.00375 to 0.12912 (Figure 1-3; MS 1-1/C 1-1 and MS 3-3/C 5-F, respectively). Between Pinaceae and Taxodiaceae, the variation among sequence divergence between all pairwise comparisons of type I, II, and type III CAD

sequences from *Metasequoia glyptostroboides* and all three *Abies* species are even more striking (Figure 1-4). The resulting sequence divergence values varied as much as 62-fold, ranging from 0.00191 to 0.11879 (Figure 1-4; AF 2/MS 1 and AB 3/MS 1-3, respectively).

DISCUSSION

Because phylogenetic analysis of the type I CAD sequences yielded a topology that is congruent with the Pinaceae 'species phylogeny', it is likely that the type I CAD sequences represent orthologous relationships at the intergeneric level. In addition, the mean synonymous and nonsynonymous substitutions of type I CAD among the genera are not significantly different from those of the 4CL gene. These results indicate that the type I CAD sequences of Pinaceae diverged at a rate similar to that observed for the 4CL gene. The similarity of sequence divergence between the two nuclear genes suggests that such rates of sequence divergence may be representative of the rate of nuclear gene divergence in Pinaceae.

The striking level of variability in the amount of sequence divergence both within Taxodiaceae (Figure 1-3), and between Pinaceae and Taxodiaceae (Figure 1-4) led us to investigate further the evolutionary dynamics of type II and type III CAD. Unlike the type I CAD sequences for Pinaceae, we do not have a 4CL data set of Taxodiaceae species for comparison to the type II and III CAD sequences, and, because the two families may not evolve at the same rate, we can not directly apply the CAD divergence rate in Pinaceae to Taxodiaceae. Therefore, to establish the expected amount of sequence divergence between orthologous CAD sequences both within Taxodiaceae, and between

Figure 1-3. Pairwise comparisons of sequence divergence between all sequences obtained from *Metasequoia glyptostroboides* (MS) and *Cryptomeria japonica* (C). Comparison categories are as follows: horizontal lines – within IIA, solid black – within IIB, diagonal lines – between IIA and IIB, solid gray – between IIA and III, dots – between IIB and III. The two horizontal dotted lines represent the expected amount of divergence ($d_{MC(CAD)}$) based on the *rbc*L (upper) and 18S (lower) estimations

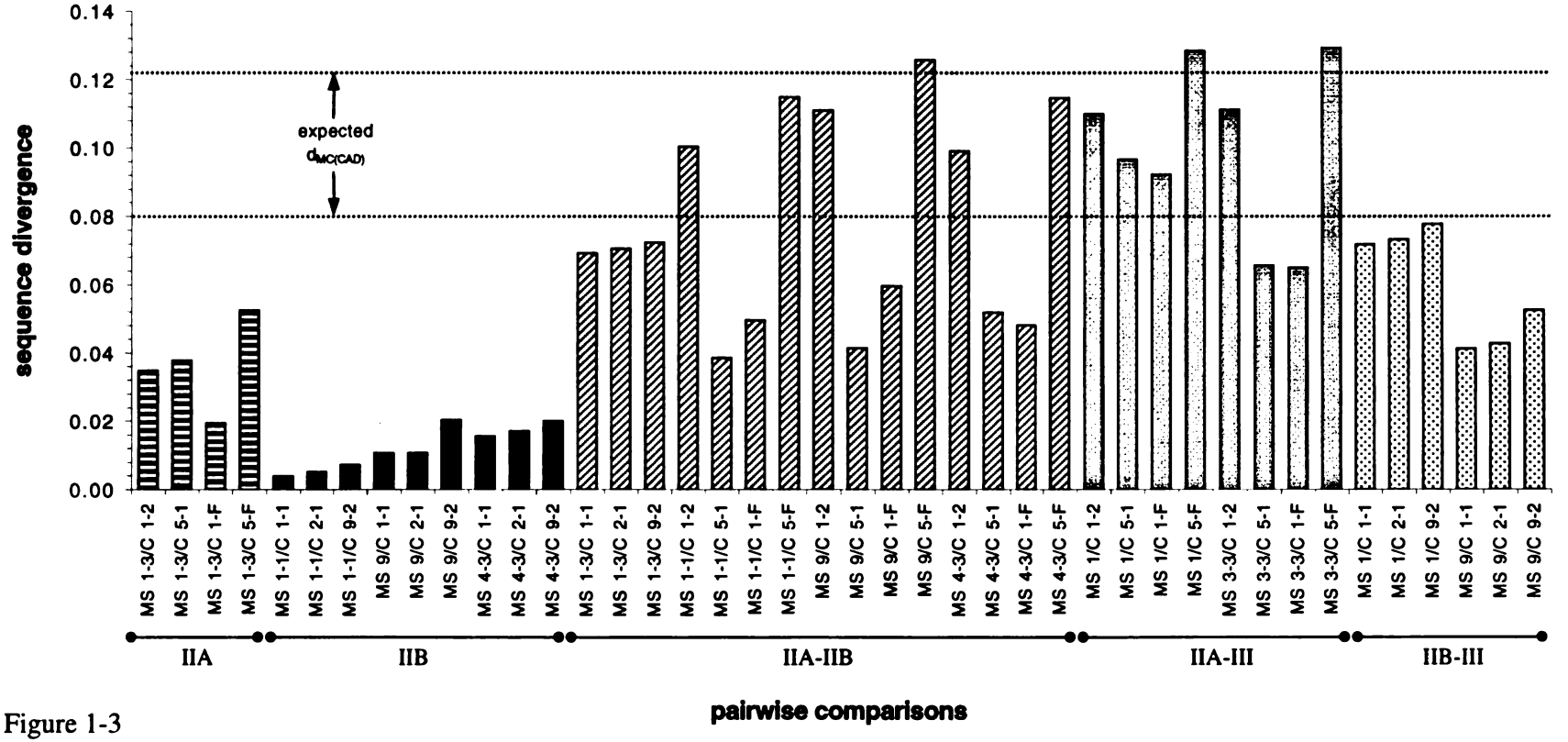

sequence divergence
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
expected
$d_{MC(CAD)}$
MS 1-3/C 1-2
MS 1-3/C 5-1
MS 1-3/C 1-F
MS 1-3/C 5-F
MS 1-1/C 1-1
MS 1-1/C 2-1
MS 1-1/C 9-2
MS 9/C 1-1
MS 9/C 2-1
MS 9/C 9-2
MS 4-3/C 1-1
MS 4-3/C 2-1
MS 4-3/C 9-2
MS 1-3/C 1-1
MS 1-3/C 2-1
MS 1-3/C 9-2
MS 1-1/C 1-2
MS 1-1/C 5-1
MS 1-1/C 1-F
MS 1-1/C 5-F
MS 9/C 1-2
MS 9/C 5-1
MS 9/C 1-F
MS 9/C 5-F
MS 4-3/C 1-2
MS 4-3/C 5-1
MS 4-3/C 1-F
MS 4-3/C 5-F
MS 1/C 1-2
MS 1/C 5-1
MS 1/C 1-F
MS 1/C 5-F
MS 3-3/C 1-2
MS 3-3/C 5-1
MS 3-3/C 1-F
MS 3-3/C 5-F
MS 1/C 1-1
MS 1/C 2-1
MS 1/C 9-2
MS 9/C 1-1
MS 9/C 2-1
MS 9/C 9-2
IIA
IIB
IIA-IIB
IIA-III
IIB-III
pairwise comparisons

Figure 1-3

Figure 1-4. Pairwise comparisons of sequence divergence between all sequences obtained from *Metasequoia glyptostroboides* (MS) and *Abies* species (A). Comparison categories are as follows: solid black – within IIB, diagonal lines – within III, horizontal lines – between IIB and III, white dots on black– between IIA and III, white dots on gray – between IIA and IIB, black dots on white – between I and IIB, checkered – between I and IIA, solid gray – between I and III. The two horizontal dotted lines represent the expected amount of divergence ($d_{MA(CAD)}$)based on the *rbc*L (upper) and 18S (lower) estimations

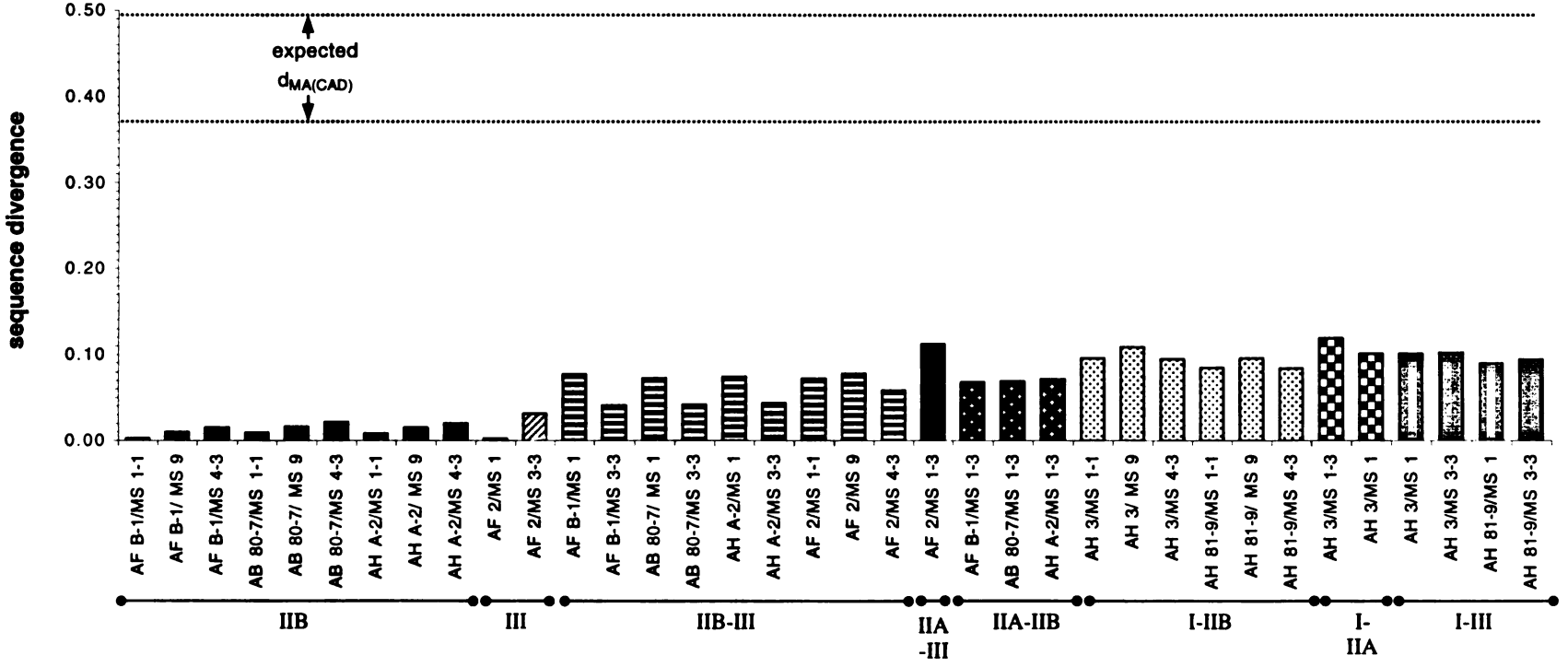

0.50
0.40
0.30
0.20
0.10
0.00
expected
$d_{MA(CAD)}$
sequence divergence
AF B-1/MS 1-1
AF B-1/ MS 9
AF B-1/MS 4-3
AB 80-7/MS 1-1
AB 80-7/ MS 9
AB 80-7/MS 4-3
AH A-2/MS 1-1
AH A-2/ MS 9
AH A-2/MS 4-3
AF 2/MS 1
AF 2/MS 3-3
AF B-1/MS 1
AF B-1/MS 3-3
AB 80-7/ MS 1
AB 80-7/MS 3-3
AH A-2/MS 1
AH A-2/MS 3-3
AF 2/MS 1-1
AF 2/MS 9
AF 2/MS 4-3
AF 2/MS 1-3
AF B-1/MS 1-3
AB 80-7/MS 1-3
AH A-2/MS 1-3
AH 3/MS 1-1
AH 3/ MS 9
AH 3/MS 4-3
AH 81-9/MS 1-1
AH 81-9/ MS 9
AH 81-9/MS 4-3
AH 3/MS 1-3
AH 3/MS 1
AH 3/MS 1
AH 3/MS 3-3
AH 81-9/MS 1
AH 81-9/MS 3-3
IIB
III
IIB-III
IIA -III
IIA-IIB
I-IIB
I- IIA
I-III
pairwise comparisons

Figure 1-4

Pinaceae and Taxodiaceae, it was necessary to estimate the ratio of sequence divergence between Taxodiaceae species and Pinaceae species in other genes.

Sequence divergence for both the cpDNA *rbc*L gene and the nuclear ribosomal 18S gene was estimated between *Metasequoia glyptostroboides* and *Cryptomeria japonica* (d_{MC}), and *Pinus wallichiana* and an *Abies* species (*rbc*L, *A. holophylla*; 18S, *A. lasiocarpa*; d_{PA}; Figure 1-5A). The ratio of divergence between the two Pinaceae species and the two Taxodiaceae species ($r_{rbcL,\ 18S} = d_{MC}/d_{PA}$) were 0.7547 and 0.5269, for the *rbc*L and 18S genes, respectively. This ratio was used to determine the proportion of sequence divergence expected between orthologous copies of CAD from *Metasequoia glyptostroboides* and *Cryptomeria japonica* ($d_{MC(CAD)}$). To accomplish this, the average sequence divergence of type I CAD between *Pinus* and *Abies holophylla* was determined ($d_{PA(CAD)} = 0.1617$), and multiplied by r_{18S} and r_{rbcL} (Figure 1-5A). The result is a window of expected sequence divergence ($d_{MC(CAD)} = 0.0852$-0.1221) for orthologous CAD sequences between *Metasequoia glyptostroboides* and *Cryptomeria japonica* (Figure 1-3; Figure 1-5A). To evaluate the amount of CAD sequence divergence between the two families, a similar analysis was conducted using type I, II, and type III CAD sequence divergences between *Metasequoia glyptostroboides* and all three *Abies* species (d_{MA}; Figure 1-5A). The ratios, r_{18S} and r_{rbcL}, were estimated to be 2.191 and 2.919, respectively, and the resulting window of expected divergence between orthologous copies of the CAD gene from *Metasequoia glyptostroboides* and *Abies* was determined to be between 0.3663 and 0.4879 (Figure 1-4; Figure 1-5A).

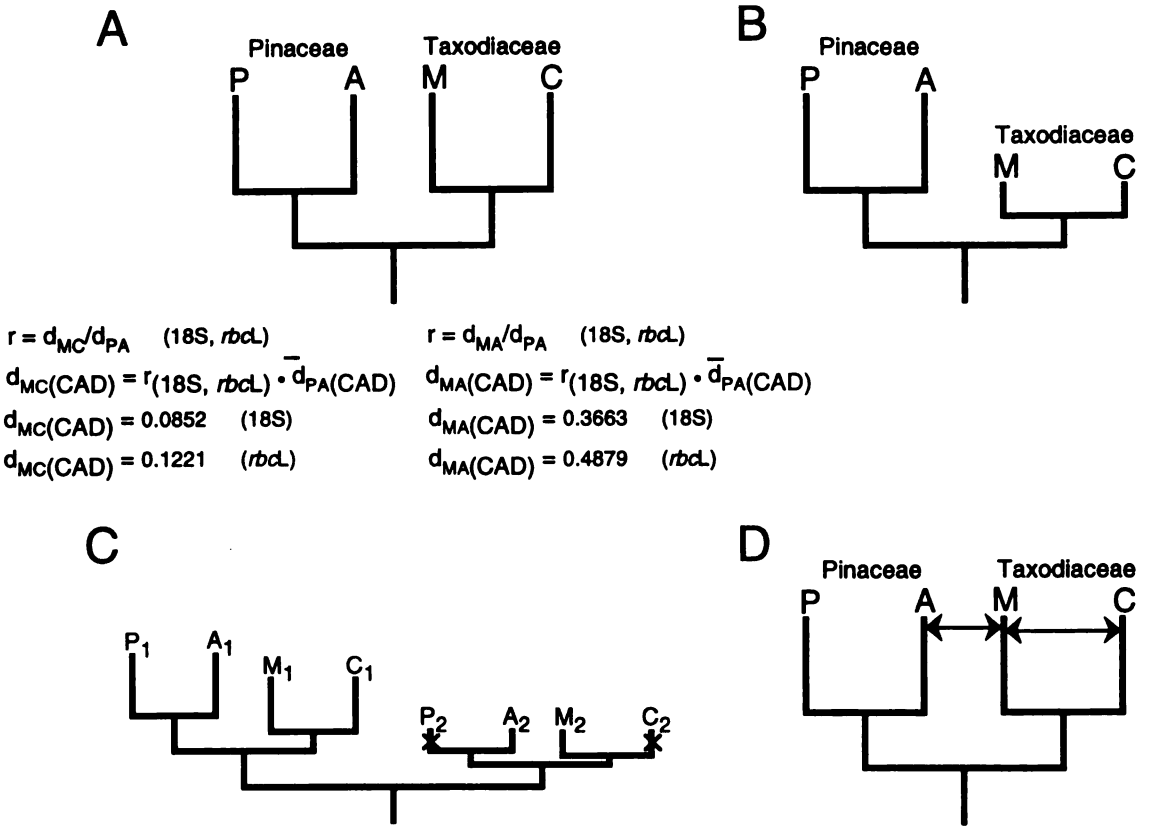


Figure 1-5. Models of divergence between *Pinus* (P), *Abies* (A), *Metasequoia* (M), and *Cryptomeria* (C): *A*, equations used for all *rbc*L and 18s approximations; d_{MC}, sequence divergence between *Metasequoia* and *Cryptomeria*; d_{PA}, sequence divergence between *Pinus* and *Abies*; $d_{PA(CAD)}$, average CAD divergence between all *Pinus* and *Abies* species; $d_{MC(CAD)}$, expected sequence divergence between orthologous CAD copies between *Metasequoia* and *Cryptomeria*; $d_{MA(CAD)}$, expected sequence divergence between orthologous CAD copies between *Metasequoia* and *Abies*; *B*, illustration of low sequence divergence in Taxodiaceae with respect to Pinaceae; *C*, illustration of the maintenance of paralogous loci between Pinaceae and Taxodiaceae, X on a branch represents a random deletion; *D*, illustration of lateral gene transfer within Taxodiaceae and between Pinaceae and Taxodiaceae

When all the pairwise comparisons of type II and type III CAD sequence divergences between *Metasequoia glyptpstroboides* and *Cryptomeria japonica* (d_{MC}) are compared to the expected amount of sequence divergence between orthologous loci of CAD for these two species, the divergence between type IIA, IIB, and type III CAD is such that some of the pairwise comparisons fall into the window of expected divergence, and thus, could reflect orthologs of the CAD gene diverging at a rate that is comparable to that seen within Pinaceae (Figure 1-3). In contrast, d_{MC} within type IIA and type IIB CAD is 1.6 – 4 and 6 – 214-times lower, respectively, than the expected amount of divergence based on the 18S and *rbc*L estimations (Figure 1-3). However, from this analysis alone it is impossible to determine which of the type II and III CAD sequences are truly orthologous.

These observations are even more extraordinary when CAD sequence divergences are examined between Pinaceae and Taxodiaceae. Figure 1-4 represents all of the pairwise comparisons of type I, II and type III CAD divergences between *Metasequoia glyptostroboides* and the three *Abies* species (d_{MA}). All of the d_{MA} values for any combination of the two species are considerably lower than that expected based on the 18S and *rbc*L approximations (Figure 1-4). Most striking are those comparisons within the type IIB and type III CAD groups. Within type IIB, d_{MA} values range from 17 – 200-times lower than the expected amount of divergence, and within the type III CAD they range from 17 – 256-times less than expected.

There are three possible explanations for the non-monophyly of type II and III CAD sequences of Pinaceae and Taxodiaceae, and the extraordinarily low type II and III CAD divergence rates observed both within Taxodiaceae, and between the two families:

1) contamination of genomic DNAs either through DNA isolation or PCR reactions, 2) extensive paralogy within and between type II and III CAD, combined with an extremely low divergence rate at some of the paralogous loci, and 3) lateral gene transfer both between genera of Taxodiaceae, and between Pinaceae and Taxodiaceae. Each of these hypotheses will be discussed in detail below.

It is very unlikely that the Pinaceae clones nested within the type II and III CAD sequences from Taxodiaceae are the result of contamination for the following reasons. First, genomic DNAs used in this study were isolated at different times, and in different laboratories, making it impossible for contamination to have taken place at the time of DNA isolation. For example, genomic DNA from the three *Abies* species (each with one or more type II and/or type III CAD clones) was isolated in the Laboratory for Systematic and Evolutionary Botany between 1996 and 1998 (Institute of Botany, the Chinese Academy of Sciences, Beijing), while genomic DNAs from the Taxodiaceae species were isolated at Michigan State University in 1998. Second, PCR contamination is unlikely because the PCR reactions for the two families were carried out at separate times, often using different combinations of the multiple CAD primers. In addition, PCR reactions were repeated multiple times for those type II and III CAD clones in which the sequence divergence and/or topological position was questionable. Third, if PCR contamination did occur, it is extremely unlikely that all three *Abies* species would become contaminated with type IIB CAD of Taxodiaceae, while none of the other Pinaceae species were. Finally, if the DNAs or PCR reactions were contaminated by Taxodiaceae species, one would expect some of the contaminated Pinaceae clones to be identical to some of the clones from Taxodiaceae species. However, none of the type IIB or type III

CAD clones isolated from Pinaceae species are identical in sequence to any of the Taxodiaceae clones. Therefore, DNA and/or PCR contamination can be ruled out as a cause of the atypical patterns observed among the type II and III CAD sequences.

The second hypothesis explaining the observed pattern of divergence of type II and III CAD relies on two mechanisms: 1) extensive paralogy within and between type II and III CAD sequences, and 2) an extremely low rate of divergence both within Taxodiaceae, and between the type II and/or type III CAD sequences from species of Pinaceae and Taxodiaceae. In contrast to the topology observed within both type II and type III CAD sequences, all previous phylogenetic hypotheses, based on both morphological and molecular data (e.g., 18S and *rbc*L), strongly support the monophyly of each of the conifer families (Chase et al. 1993; Brunsfeld et al. 1994; Tsumura et al. 1995). Similarly, the topological relationships within Taxodiaceae observed on the CAD gene tree (Figure 1-2) are incongruent with that of the *rbc*L phylogeny (Brunsfeld et al. 1994). Topological discordance between gene trees and the corresponding species tree could be a result of sampling paralogous loci between taxa (Doyle 1992; Maddison 1997; Page and Charleston 1997). The observed pattern of divergence among type II and III CAD sequences (Figure 1-2) could have resulted from sampling paralogous CAD loci, both within Taxodiaceae, and between Taxodiaceae and Pinaceae. For example, if a duplication of the CAD gene occurred before the diversification of the two families, and, following diversification, paralogous CAD loci were randomly deleted from each of the two families, the resulting topology would reveal a pattern of gene evolution that is different than the species phylogeny (Figure 1-5C). However, this is only a simple example illustrating the mechanism by which this discordance could have been

established. To explain the observed topology of type II and type III CAD sequences, duplication and deletion of the CAD gene would have had to occur extensively throughout the evolution of the two conifer families. To assure that all loci present in each species were isolated, we screened each using multiple type specific CAD primers (Figure 1-1, Table 1-2). Therefore, it is unlikely that loci not identified in some species were not sampled by PCR, but rather, these loci were randomly deleted.

In addition to the extensive duplication and deletion, this hypothesis relies on there also being an extremely low rate of divergence at some of the duplicate loci, both within Taxodiaceae and between the two families. The divergence between *Metasequoia glyptostroboides* and *Cryptomeria japonica* (Figure 1-3) is as much as 214-times lower than expected. Therefore, for any orthologous relationships to exist within the type IIA or type IIB sequences, the divergence rate of CAD between Taxodiaceae species must be much slower than that observed between the orthologous type I CAD sequences from Pinaceae (Figure 1-5B). Likewise, the rate of divergence of type II and/or type III CAD sequences between Pinaceae and Taxodiaceae species must be even slower to result in the strikingly low sequence divergences (as much as 256-times lower than expected) that were observed between *Metasequoia glyptostroboides* and the three *Abies* species (Figure 1-4).

Taxodiaceae is one of the oldest families of conifers, and is complimented by an extensive fossil record for most of the genera (Florin 1963). Even the most closely related genera (*Metasequoia*, *Sequoia*, and *Sequoiadendron*) have been separated for at least 100 million years, as they all are present in the fossil record from the late Cretaceous. The original diversification of the family likely occurred in the Jurassic

(>180 mya), as fossil evidence suggests that the *Cryptomeria*-like genus *Sewardiodendron* was established at this time (Yao, Zhou, and Zhang 1998). Like Taxodiaceae, Pinaceae has an excellent fossil record, and was well established by the early Cretaceous (~140 mya; Florin 1963). The fossil record suggests that Pinaceae likely diversified sometime in the Jurassic period, and that the two conifer families have been separated for at least 200 million years. Thus, the observed divergences of the type II and III CAD sequences both within Taxodiaceae, and between the two families, is much lower than expected for this amount of time. For example, using 200 million years as the time of divergence between the two families, the substitution rate between the type III CAD clones *Abies firma* 2 and *Metasequoia glyptostroboides* 1 (sequence divergence = 0.00191; Figure 1-2) would be only 9.6×10^{-12} substitutions per site per year. This extremely low substitution rate at all sites (including nonsynonymous sites and introns) between the two CAD clones is nearly 600-times less than a previous estimate of synonymous substitution rate in plant nuclear genes of $4.1\text{-}5.7 \times 10^{-9}$ substitutions per site per year (Li 1997). There is no reported mechanism that can explain how such striking sequence similarity can be maintained for such a long period of time.

The third hypothesis explaining the evolution of type II and type III CAD in the two families is that of lateral gene transfer both within Taxodiaceae, and between the two families. This hypothesis assumes that the strikingly low sequence divergence values are due to the relatively recent horizontal movement of type II and III CAD within and between the two families (Figure 1-5D). To explain all type II and/or type III CAD identified in Pinaceae species, there must have been multiple lateral gene transfer events that have occurred repeatedly and independently.

It is likely that there were at least three lateral transfer events between the two families corresponding to the Pinaceae sequences nested within the type IIA and IIB CAD, and the *Metasequoia glyptostroboides* type III CAD sequences (Figure 1-2). Because both type IIA and IIB CAD sequences are dominated by Taxodiaceae species, it is most parsimonious to assume that the transfer occurred from the Taxodiaceae species to the species of Pinaceae nested within the group. For example, one type IIB CAD clone was isolated from each of the three *Abies* species included in the analysis. These sequences probably represent one transfer event from a Taxodiaceae species to *Abies* before the diversification of the three *Abies* species. In the type III CAD group, since *Metasequoia glyptostroboides* is the only Taxodiaceae species with this type of CAD sequence, it is most parsimonious to assume that the type III CAD donor was Pinaceae, rather than Taxodiaceae. Similarly, to explain the extremely low CAD divergence values observed within Taxodiaceae, in addition to those between the two families, there must have been multiple lateral gene transfer events within Taxodiaceae as well.

Most previous documentation of lateral gene transfer has been from bacteria and fungi (e.g., Nelson et al. 1999; Screen and St Leger 2000), and between plants and associated bacteria (e.g., Aoki and Syono 1999). Between higher plants, previous hypotheses of lateral gene transfer have been limited to a mobile group I intron found in the mitochondrial *cox1* gene (Cho et al. 1998; Cho and Palmer 1999). If lateral gene transfer is the case for the CAD gene here, this would be the first documentation of the transfer of a structural, protein-coding gene laterally between plant species.

CONCLUSIONS

The two competing hypotheses are both unique, as neither of these phenomena have been previously reported in plants, making it difficult to speculate which of the two is more likely. However, unlike lateral gene transfer, sequence divergence has been studied extensively in a large number of genes from a diverse assemblage of plants. In addition, the mechanisms influencing sequence divergence are much more clearly understood than those of lateral gene transfer. To our knowledge, there is no known evolutionary mechanism that can explain the maintenance of this level of sequence similarity between Pinaceae and Taxodiaceae, or even within Taxodiaceae, over the 200 million-year history of the two families. However, few studies have focussed on the evolution of nuclear genes in conifers. Nevertheless, based on the analyses presented in this study, it is very unlikely that the relationships portrayed in the NJ tree (Figure 1-2) could have resulted through paralogy and low sequence divergence alone. Therefore, although lateral gene transfer is a poorly understood phenomenon, it is more likely that the observed patterns of divergence are the result of the movement of genes horizontally between species, via an insect and/or pathogen (fungal or bacterial) vector. The overall complexity of the conifer nuclear genome is most likely the result of an amalgam of evolutionary mechanisms like those hypothesized here. However, our understanding of molecular evolution in conifers is still in its infancy, and there is a need for continued research to this end in the future.

LITERATURE CITED

AOKI, S.,and K. SYONO. 1999. Horizontal gene transfer and mutation: the *Ngrol* genes in the genome of *Nicotiana glauca*. Proc. Natl. Acad. Sci. USA **96**: 13229-13234.

BRUSFELD, S. J., P. S. SOLTIS, D. E. SOLTIS, P. A. GADEK, C. J. QUINN, D. D. STRENGE, and T. M. RANKER. 1994. Phylogenetic relationships among the genera of Taxodiaceae and Cupressaceae: evidence from *rbc*L sequences. Syst. Bot. **19**: 253-262.

CHASE, M. W., D. E. SOLTIS, R. G. OLMSTEAD, et al. 1993. Phylogenetics of seed plants - an analysis of nucleotide-sequences from the plastid gene *rbc*L. Ann. Mo. Bot. Gard. **80**: 528-580.

CHO, Y., Y. L. QIU, P. KUHLMAN, and J. D. PALMER. 1998. Explosive invasion of plant mitochondria by a group I intron. Proc. Natl. Acad. Sci. USA **95**: 14244-14249.

CHO, Y. R. and J. D. PALMER. 1999. Multiple acquisitions via horizontal transfer of a group I intron in the mitochondrial cox1 gene during evolution of the Araceae family. Mol. Biol. Evol. **16**: 1155-1165.

DOYLE, J. J. and J. L. DOYLE. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem. Bull. **19**: 11-15.

DOYLE, J. J. 1992. Gene trees and species trees – molecular systematics as one-character taxonomy. Syst. Bot. **17**: 144-163.

DOYLE, J. J., V. KANAZIN, and R. C. SHOEMAKER. 1996. Phylogenetic utility of histone H3 intron sequences in the perennial relatives of soybean (*Glycine*: Leguminosae). Mol. Phylog. Evol. **6**: 438-447.

EMSHWILLER, E. and J. J. DOYLE. 1999. Chloroplast-expressed glutamine synthetase (ncpGS): potential utility for phylogenetic studies with an example from *Oxalis* (Oxalidaceae). Mol. Phylog. Evol. **12**: 310-319.

FARJON, A. 1998. World Checklist and Bibliography of Conifers, Royal Botanic Gardens, Kew, UK.

FLORIN, R. 1963. The distribution of conifer and taxad genera in time and space. *Acta Hort. Berg.* **20**: 121-312.

GOTTLIEB, L. D. and V. S. FORD. 1996. Phylogenetic relationships among the sections of *Clarkia* (Onagraceae) inferred from the nucleotide sequences of *PgiC*. Syst. Bot. **21**: 45-62.

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

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

KINLAW, C. S. and D. B. NEALE. 1997. Complex gene families in pine genomes. Trends Plant Sci. **2**: 356-359.

KISHINO, H., and M. HASEGAWA. 1989. Evaluation of the maximum likelihood estimates of the evolutionary tree topologies from sequence data, and the branching order in Hominoidea. J. Mol. Evol. **29**: 170-179.

KUMAR, S., K. TAMURA, and M. NEI. 1993. MEGA: molecular evolutionary genetics analysis. Version 1.02. The Pennsylvania State University, University Park, PA.

LI, W.-H. 1997. Molecular Evolution. Sinauer Associates, Sunderland, MA.

MACKAY, J. J., W. LIU, R. WHETTEN, R. R. SEDEROFF, and D. M. O'MALLEY. 1995. Genetic analysis of cinnamyl alcohol dehydrogenase in loblolly pine: single gene inheritance, molecular characterization and evolution. Mol. Gen. Genet. **247**: 537-545.

MACKAY, J. J., D. M. O'MALLEY, T. PRESNELL, F. L. BOOKER, M. M. CAMPBELL, R. W. WHETTEN, and R. R. SEDEROFF. 1997. Inheritance, gene expression, and lignin characterization in a mutant pine deficient in cinnamyl alcohol dehydrogenase. Proc. Natl. Acad. Sci. USA **94**: 8255-8260.

MADDISON, W. P. 1997. Gene trees in species trees. Syst. Biol. **46**: 523-536.

MASON-GAMER, R. J., C. F. WEIL, and E. A. KELLOGG. 1998. Granule-bound starch synthase: Structure, function, and phylogenetic utility. Mol. Biol. Evol. **15**: 1658-1673.

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

MILLER, C. N. 1977. Mesozoic conifers. Bot. Rev. **43**: 217-281.

MOYLE, R., A. WAGNER, and C. WALTER. 1998. Nucleotide sequence of a cinnamyl alcohol dehydrogenase gene (accession no. AF060491) from *Pinus radiata* (PGR98-118). Plant Physiol. **117**: 1125.

MURRAY, B. G. 1998. Nuclear DNA amounts in gymnosperms. Ann. Bot. **82** (Supplement A): 3-15.

NELSON, K. E., R. A. CLAYTON, S. R. GILL, et al. 1999. Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermatoga maritima*. Nature **399**: 323-329.

O'MALLEY, D.M., S. PORTER, and R. R. SEDEOFF. 1992. Purification, characterization, and cloning of cinnamyl alcohol dehydrogenase in loblolly pine (*Pinus taeda* L.). Plant Physiol. **98**: 1364-1371.

PAGE, R. D. M. and M. A. CHARLESTON. 1997. From gene to organismal phylogeny: reconciled trees and the gene tree species tree problem. Mol. Phylog. Evol. **7**: 231-240.

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

SANG, T., DONOGHUE, M. J., and D. ZHANG. 1997. Evolution of alcohol dehydrogenase genes in peonies (*Paeonia*): phylogenetic relationships of putative nonhybrid species. Mol. Biol. Evol. **14**: 994-1007.

SCHUBERT, R., C. SPERISEN. G. MUELLER-STARCK, S. LA SCALA, D. ERNST, H. SANDERMAN JR., and K. P. HAEGER. 1998. The cinnamyl alcohol dehydrogenase gene structure in *Picea abies* (L.) Karst. : genomic sequences, Southern hybridization, genetic analysis and phylogenetic relationships. Trees **12**: 453-463.

SCREEN, S. E., and R. J. ST LEGER. 2000. Cloning, expression, and substrate specificity of a fungal chymotrypsin – evidence for lateral gene transfer from an actinomycete bacterium. J. Biol. Chem. **275**: 6689-6694.

SMALL, R. L., J. A. RYBURN, R. C. CRONN, T. SEELANAN, and J. F. WENDEL. 1998. The tortoise and the hare: choosing between noncoding plastome and nuclear *Adh* sequences for phylogeny reconstruction in a recently diverged plant group. Amer. J. Bot. **85**: 1301-1315.

SWOFFORD, D. L. 1998. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other methods). Version 4. Sinauer Associates, Sunderland, MA.

TAMURA, K. and M. NEI. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial-DNA in humans and chimpanzees. Mol. Biol. Evol. **10**: 512-526.

TEMPLETON, A. R. 1983. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. Evolution **37**: 221-244.

THOMPSON, J. D., D. G. HIGGINS, and T. J. GIBSON. 1994. CLUSTAL-W - improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**: 4673-4680.

TSUMURA, Y., K. YOSHIMURA, N. TOMARU, and K. OHBA. 1995. Molecular phylogeny of conifers using RFLP analysis of PCR-amplified specific chloroplast genes. Theor. Appl. Genet. **91**: 1222-1236.

WALTER, M. H., J. GRIMA-PETTENATI, C. GRAND, A. M. BOUDET, and C. J. LAMB. 1988. Cinnamyl-alcohol dehydrogenase, a molecular marker specific for lignin biosynthesis: cDNA cloning and mRNA induction by a fungal elicitor. Proc. Natl. Acad. Sci. USA. **86**: 5546-5550.

WANG, X.-Q., D. C. TANK, T. SANG. 2000. Phylogeny and divergence times in Pinaceae: evidence from three genomes. Mol. Biol. Evol. **17**: 773-778.

YANG, Z. B. 1994. Estimating the pattern of nucleotide substitution. J. Mol. Evol. **39**: 105-111.

YAO, X. L., Z. Y. ZHOU, and B. L. ZHANG. 1998. Reconstruction of the Jurassic conifer *Sewardiodendron laxum* (Taxodiaceae). Amer. J. Bot. **85**: 1289-1300.

CHAPTER 2

EVOLUTION OF THE GLYCEROL-3-PHOSPHATE ACYLTRANSFERASE GENE AND ITS PHYLOGENETIC IMPLICATIONS IN *PAEONIA* (PAEONIACEAE)

INTRODUCTION

Low-copy nuclear genes have the potential to provide an abundance of independent gene phylogenies. Aside from the sheer number of potential independent markers, low-copy nuclear genes are biparentally inherited, and generally diverge at a higher rate than cpDNA or nrDNA, most notably in intron regions. Therefore, low-copy nuclear genes are especially useful for reconstructing low-level taxonomic relationships in which cpDNA and nrDNA nucleotide sequences are too conserved to resolve.

The use of low-copy nuclear genes in molecular phylogenetic studies of plants is increasing (e.g., Gottlieb and Ford 1996; Doyle, Kanazin, and Shoemaker 1996; Sang, Donoghue, and Zhang 1997; Mason-Gamer, Weil, and Kellogg 1998; Small et al. 1998; Emshwiller and Doyle 1999; Matthews and Donoghue 1999; Wang, Tank, and Sang 2000). However, in comparison to more commonly used molecular markers in plant systematic studies (i.e., cpDNA and nrDNA genes and spacers), the phylogenetic utility of low-copy nuclear genes is still largely understudied. This is due primarily to difficulties in determining orthology from paralogy among members of a gene family, and the increased lab-work necessary for cloning. The selection of genes that exist in

relatively small gene families, and are less dynamic in duplication and deletion can aid in overcoming these difficulties.

In this study we investigated the phylogenetic utility of the chloroplast-expressed glycerol-3-phosphate acyltransferase (GPAT) nuclear gene in determining interspecific relationships in the angiosperm genus *Paeonia* (Paeoniaceae). GPAT is an essential enzyme utilized in the catalysis of the initial step of glycerolipid synthesis, specifically, the formation of lysophosphotidic acid from glycerol-3-phosphate and acylthioesters, in the cells of all higher organisms (Nishida et al. 1993). The GPAT gene was selected for this study because it is well studied in angiosperms, with sequences available in GenBank from five different families, including Brassicaceae (Nishida et al. 1993), Fabaceae (Weber et al. 1991), Asteraceae (Bhella and MacKenzie 1994), Chenopodiaceae (Wolter, unpublished), and Cucurbitaceae (Ishizaki et al. 1988). Based on enzyme activity in mutants of *A. thaliana* and Southern blot hybridizations, the GPAT gene has been characterized as single copy in all five angiosperm families. The presence of the GPAT gene at a single locus in the five eudicot families suggests that the GPAT gene may not be subject to dynamic cycles of duplication and deletion. Therefore, this gene has the potential to serve as a useful marker for phylogenetic studies of angiosperms.

Paeonia is classified in the monogeneric family Paeoniaceae, and is comprised of ~35 species of herbaceous and woody habit disjunctly distributed in five areas of the Northern Hemisphere. *Paeonia* has been further divided into three sections, *Moutan, Onaepia,* and *Paeonia*. The largest section, *Paeonia*, contains ~28 herbaceous diploid and tetraploid species distributed in eastern and central Asia, the western Himalayas and the European Mediterranean region. Based on leaf morphology, section *Paeonia* has

been further partitioned into two subsections, *Paeonia* and *Foliolatae*. Section *Moutan* is comprised of five diploid woody species, in two subsections *Delavayanae* and *Vaginitae*, distributed in central and western China. The smallest section, *Onaepia*, consists of only two diploid herbaceous species endemic to Pacific North America (Stern 1946; Pan 1979; Tzanoudakis 1983; Pei 1993).

Previous phylogenetic hypotheses based on nucleotide sequences from multiple genic and intergenic regions, including, two loci of the low-copy nuclear gene alcohol dehydrogenase (*Adh*), *Adh1* and *Adh2*, the cpDNA gene *mat*K and two intergenic spacers *trn*L-*trn*F and *psb*A-*trn*H, and the nrDNA ITS region, support the monophyly of each of the three sections of *Paeonia*, as well as each subsection of section *Moutan*. *Adh* gene phylogenies also support the sister relationship of section *Paeonia* and section *Onaepia*. In addition, these analyses have indicated a complex pattern of reticulate evolution within section *Paeonia* (Sang, Crawford, and Stuessy 1995, 1997; Sang, Donoghue, and Zhang 1997; Sang and Zhang 1999).

The primary objectives of this study were to 1) investigate the molecular evolution of the chloroplast-expressed GPAT gene in *Paeonia* through comparison to previous phylogenetic hypotheses, and 2) investigate the phylogenetic utility of the GPAT gene in *Paeonia* in hopes to develop an additional independent nuclear gene marker for studying the complex phylogenetic relationships in *Paeonia*.

MATERIALS AND METHODS

Sampling of *Paeonia* species for this investigation included 13 species representing each of the three sections. From section *Paeonia*, 7 species were sampled, including, the four diploid members of subsection *Paeonia*, *P. anomola*, *P. lactiflora*, *P.*

tenuifolia, and *P. veitchii* (2 populations), and 3 species from subsection *Foliolatae*, the diploid *P. japonica*, three populations of both diploid and tetraploid *P. obovata*, and the tetraploid *P. mairei*. We sampled all five species of section *Moutan*; *P. delavayi*, *P. lutea* (two populations), *P. rockii* (two populations), *P. suffruticosa* ssp. *spontanea*, and *P. szechuanica*. To represent section *Onaepia*, two populations of *P. californica* were sampled. Total DNAs were isolated previously using the CTAB method (Doyle and Doyle 1987) from leaves of *Paeonia* species collected from natural populations in Europe, China and the United States (Sang, Crawford, and Stuessy 1997). Additional sampling for this study includes two new populations of *P. obovata* (*P. obovata-2*, and *P. obovata-3*) collected from natural populations in China.

PCR AND SEQUENCING

Two general PCR primers, GAF1 (5' – TTTGGYCAAAATTATATTCGKCC) and GAR1 (5' – CCACCACTKGGTGCAATCCA; Figure 2-1), were designed in the most conserved regions across GPAT sequences from five eudicot families, including Brassicaceae (Nishida et al. 1993), Fabaceae (Weber et al. 1991), Asteraceae (Bhella and MacKenzie 1994), Chenopodiaceae (Wolter, unpublished), and Cucurbitaceae (Ishizaki et al. 1988). PCR amplification using the general primers yielded an approximately 300 bp fragment from *P. californica*, which upon sequencing was determined to be a retrogene containing only exon sequences. Due to the presence of two very large introns in all *Paeonia* species downstream of the GAF1 primer, we were unable to amplify this portion of the gene from other *Paeonia* species. As a result, two peony specific PCR primers, GAF2 (5' – AGCAGACCCTGCTATCATTGC) and GAR2 (5' –

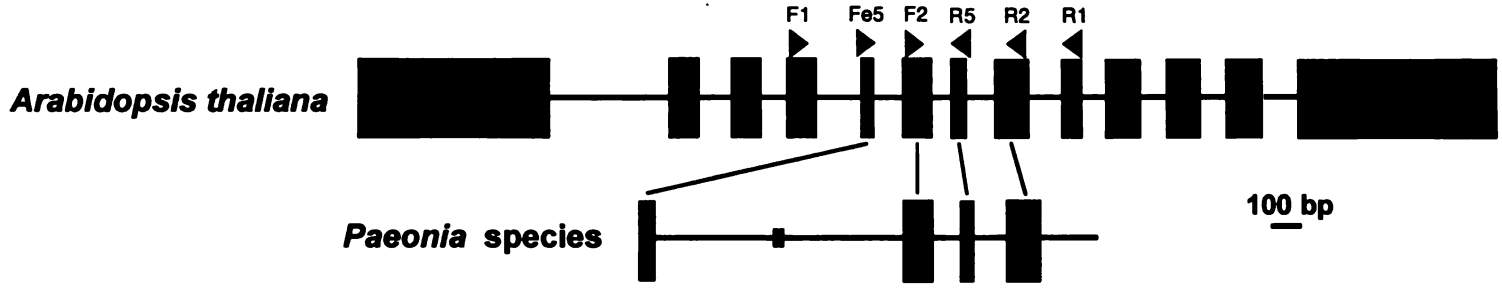


Figure 2-1. Diagram of the full-length GPAT gene in *Arabidopsis thaliana* and a portion of the gene in peonies. Boxes represent exons, and lines between exons represent introns. Lines connecting exons between *A. thaliana* and *Paeonia* species indicate homologous exons. Arrows above exons indicate the location and direction of PCR primers used in this study. The size of each region is measured by the scale bar except where indicated

TCAGCAAGCTCAGGAACATCA; Figure 2-1), were designed based on nucleotide sequence of the *P. californica* retrogene.

Preliminary phylogenetic analyses of an ~580 bp portion of the GPAT gene amplified with the primers GAF2 and GAR1 from a number of *Paeonia* species representing all three sections indicated the potential phylogenetic utility the GPAT gene in *Paeonia*. However, because of the relatively short fragment of the GPAT gene used in these analyses, many of the interspecific relationships in the genus remained unresolved. Therefore, to design PCR primers to amplify a larger portion of the gene, we screened a previously constructed genomic library from *P. anomola* in an attempt to isolate a genomic clone containing the full-length GPAT gene (see below).

For all PCR amplifications in this study, the GPAT gene was amplified through the following PCR cycles: (1) 70°C, 4 min; (2-4) 94°C, 1 min; 52-55°C, 30 sec; 72°C, 2 min; (5-7) 94°C, 20 sec; 52-55°C, 30 sec; 72°C, 2 min; (8) repeat steps 5-7 29 times; (9) 72°C, 10 min. All resulting PCR products were cloned with a Topo-TA™ cloning kit (Invitrogen). For each species, 10 to 20 clones were screened by examining restriction-site or sequence (from one primer) variation (Sang, Donoghue, and Zhang 1997; Wang, Tank, and Sang 2000). Distinct clones were fully sequenced and included in the phylogenetic analyses. Sequencing was done on an ABI 373 automated DNA sequencer using either the Dye Terminator Cycle Sequencing reaction kit (PE Applied Biosystems) or the DYEnamic ET Terminator Cycle Sequencing reaction kit (Amersham Pharmacia Biotech). Upon submission for publication, all sequences will be deposited in GenBank.

GENOMIC LIBRARY SCREENING

The genomic library was screened with a ^{32}P-labeled probe constructed by random priming with the GAF2/GAR1 GPAT fragment from *P. veitchii* following protocols of Sambrook, Fritsch, and Maniatas (1989). Two positive clones (C2 and C3) were isolated and purified with a Qiagen Lambda DNA Mini Kit (Qiagen Inc.), analyzed by restriction digestion, and determined to be identical. The genomic clone C3 was characterized by restriction mapping (Figure 2-2A), and subcloned using a Zero Background™/Kan Cloning Kit (Invitrogen). Subcloned fragments were sequenced using the M13 forward and reverse sequencing primers located on the plasmid vector. To locate the GPAT gene within the C3 genomic clone and determine sequence similarity to other sequences in GenBank, BLAST searches were performed. A second round of genomic library screening using a ^{32}P-labeled probe constructed by random priming of a GPAT fragment from *P. veitchii* amplified with the primers GAF2 and a newly designed peony specific primer GAR5 (5' - CATGCTGAATGGCTTGCAAAG; Figure 2-1). One additional distinct genomic clone (C7) was isolated and purified, characterized by restriction mapping, subcloned, and sequenced as described above (Figure 2-2B). Based on sequences obtained from the C7 genomic clone, one additional PCR primer was designed (GAFe5, 5' – CCCTGTTCTCTGGAATGGAAG; Figure 2-1) to amplify a larger portion of the GPAT gene from all *Paeonia* species included in the phylogenetic analyses.

PHYLOGENETIC ANALYSES

Sequence alignments of distinct GPAT clones were performed manually. A few regions in the GPAT introns that could not be unambiguously aligned were excluded from phylogenetic analyses. Parsimony, as executed in PAUP* 4.0 (Swofford 1998), was

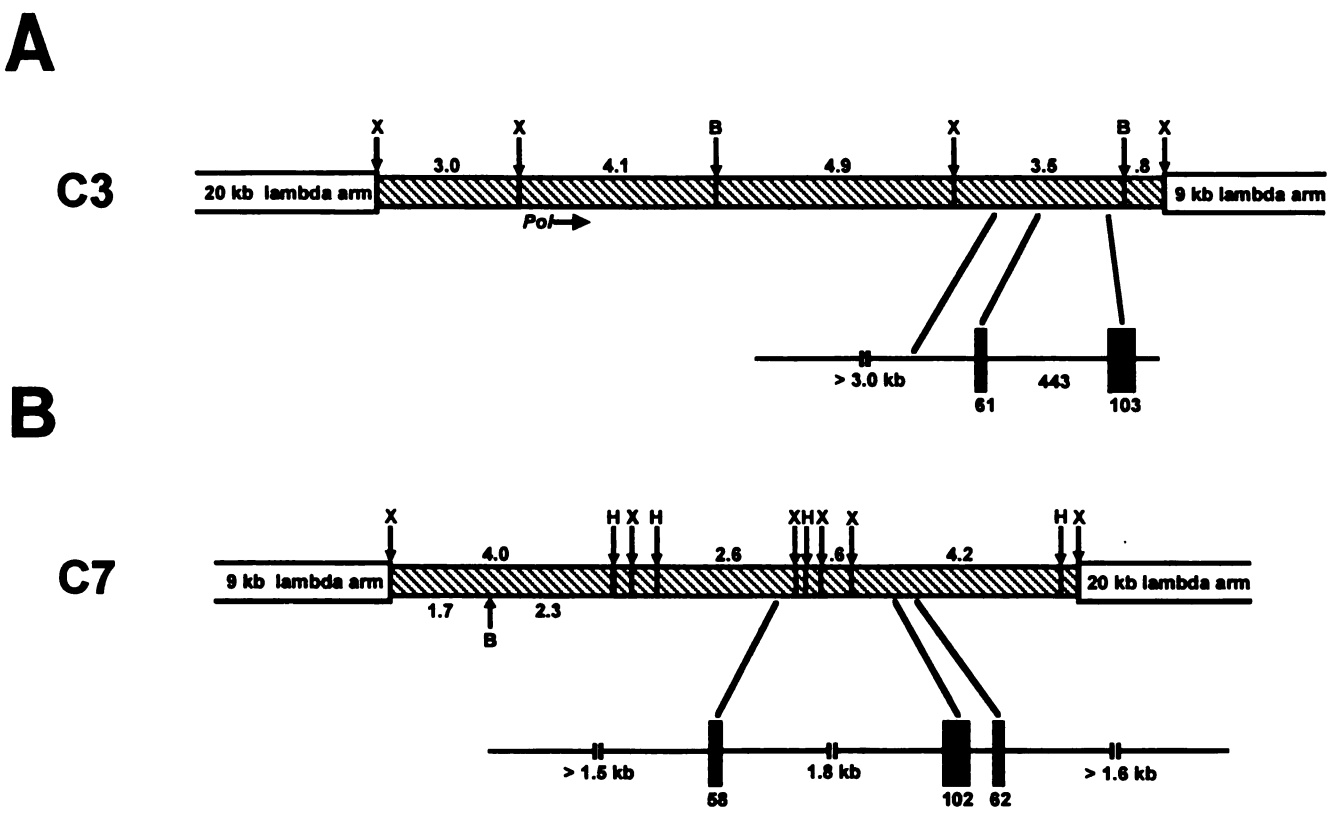


Figure 2-2. Characterization of two genomic clones isolated from *Paeonia anomala* genomic library screening. Arrows indicate the position of restriction endonuclease cut sites (B, *Bam*HI; H, *Hind*III; X, *Xba*I) used for restriction mapping and subcloning. Sizes of the resulting fragments are given in kilobases. Underneath each genomic clone characterization is blow-up of the portion of the GPAT gene identified in each with sizes given in base pairs, and lines indicating the corresponding region of the genomic clone from which they were identified. *A*, genomic clone C3, *Pol*→ indicates the position and orientation of the *Pol* gene; *B*, genomic clone C7

utilized to infer the gene phylogeny based on nucleotide substitutions in aligned partial GPAT sequences obtained from PCR amplification with the primers GAFe5 and GAR5 (Figure 2-1). Unweighted parsimony analysis was performed by heuristic search with tree bisection-reconnection (TBR) branch swapping, the MULTREES option, ACCTRAN optimization, and 10,000 random-addition replicates. Bootstrap analysis was carried out with 10,000 replicates of heuristic search with TBR branch swapping, ACCTRAN optimization, and simple taxon addition. Section *Moutan* was used as a functional outgroup based on previous phylogenetic hypotheses (Sang, Donoghue, and Zhang 1997). Topological congruence to previous phylogenetic hypotheses was assessed with the Templeton test (Templeton 1983), as implemented in PAUP* 4.0, using the *Paeonia* 'species phylogeny' as a topological constraint.

RESULTS

PCR, SEQUENCING, AND PHYLOGENETIC ANALYSES

After screening 10-20 GPAT clones from each of the 13 *Paeonia* species, only one type of clone was isolated from *P. lactiflora*, *P. obovata-1*, *P. obovata-3*, *P. rockii-2*, and *P. veitchii-4*. Two to four types of GPAT clones were isolated from all other species, and/or populations, sampled for this study. In total, 41 distinct GPAT clones were included in the phylogenetic analysis. The resulting data set contained 2,674 bp of alignable sequence, spanning three exons and two introns (Figure 2-1), of which 141 bp were of exon regions and 2,533 bp of alignable intron regions. Of the 2,533 bp of alignable intron sequence, 2,438 bp represented one large intron present in all species of *Paeonia* screened. Parsimony analysis resulted in 45 most parsimonious trees (tree length = 530, consistency index = 0.88, retention index = 0.96). One of the 45 most

parsimonious trees was randomly selected to represent the GPAT gene phylogeny with nodes that collapse on the strict consensus indicated (Figure 2-3).

Topological incongruence between the GPAT tree and previous phylogenetic hypotheses involve the sister relationship of two diploids, *P. anomola* and *P. veitchii* in subsection *Paeonia* of section *Paeonia*, and the sister relationship of *P. californica* of section *Onaepia* and subsection *Foliolatae* of section *Paeonia*. Previous phylogenetic hypotheses identified *P. anomala* as the sister species of *P. tenuifolia*, and strongly support the monophyly of both section *Paeonia* and section *Onaepia* (Sang, Crawford, and Stuessy 1995, 1997; Sang, Donoghue, and Zhang 1997; Figure 2-3). The Templeton test was performed on the GPAT phylogeny separately for each of these incongruencies, while previous phylogenetic relationships were used as a topological constraint. These analyses indicate that the relationships within section *Paeonia*, subsection *Paeonia* are not significantly incongruent ($p = 0.51$), but the incongruent sister relationship of *P. californica* (section *Onaepia*) and section *Paeonia*, subsection *Foliolatae* is highly significant ($p < 0.0001$).

GENOMIC LIBRARY SCREENING

Genomic library screening isolated two distinct genomic clones containing GPAT genes in *P. anomala* (C3 and C7; Figure 2-2). The C3 genomic clone (Figure 2-2A) was found to contain a portion of the GPAT gene including two exons and their flanking intron regions. Upon comparison to sequences obtained from other *Paeonia* species, the GPAT copy identified in genomic clone C3 was determined to be a pseudogene based on multiple insertions and deletions, as well as numerous substitutions leading to stop codons within each of the two exon regions. In addition, the intron structure of this

Figure 2-3. Phylogeny of the GPAT gene of *Paeonia*. One randomly selected tree of 45 most parsimonious trees (tree length = 530, consistency index = 0.88, retention index = 0.96). Species represented by more than one population are indicated with hyphenated population numbers following the name. Numbers following a species name indicate clone numbers. Numbers associated with the branches are bootstrap percentages greater than 50%. * = branch collapses on the strict consensus. Branch lengths are proportional to the numbers of nucleotide substitutions and are measured by the scale bar.

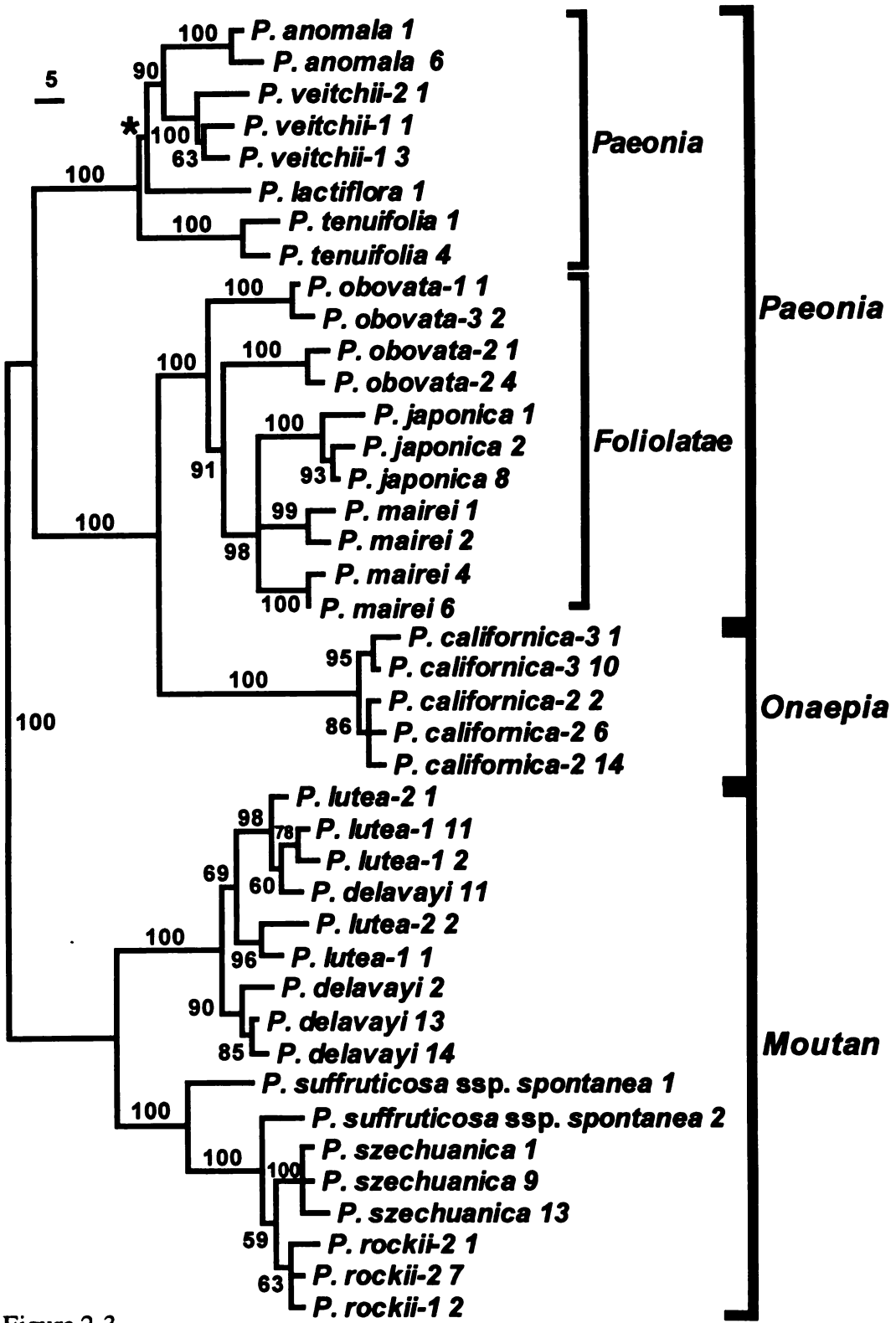

5
P. anomala 1
P. anomala 6
P. veitchii-2 1
P. veitchii-1 1
P. veitchii-1 3
P. lactiflora 1
P. tenuifolia 1
P. tenuifolia 4
P. obovata-1 1
P. obovata-3 2
P. obovata-2 1
P. obovata-2 4
P. japonica 1
P. japonica 2
P. japonica 8
P. mairei 1
P. mairei 2
P. mairei 4
P. mairei 6
P. californica-3 1
P. californica-3 10
P. californica-2 2
P. californica-2 6
P. californica-2 14
P. lutea-2 1
P. lutea-1 11
P. lutea-1 2
P. delavayi 11
P. lutea-2 2
P. lutea-1 1
P. delavayi 2
P. delavayi 13
P. delavayi 14
P. suffruticosa ssp. spontanea 1
P. suffruticosa ssp. spontanea 2
P. szechuanica 1
P. szechuanica 9
P. szechuanica 13
P. rockii-2 1
P. rockii-2 7
P. rockii-1 2
Paeonia
Foliolatae
Paeonia
Onaepia
Moutan

Figure 2-3

GPAT copy is altered, with a large insertion in the intron between the two recovered exons, and, after sequencing nearly three kb upstream of the identified GPAT region, we were unable to locate the next exon. Based on BLAST sequence similarity, another region upstream of the identified GPAT region was characterized with high identity to the *Pol* gene, a gene involved in replication and transposition of retrotransposable elements (Li 1997).

The C7 genomic clone (Figure 2-2B), isolated in the second round of genomic library screening probed with a different GPAT region (see above), was determined to contain a functional copy of the GPAT gene in *P. anomola*. Based on sequence comparisons to the two GPAT clones isolated via PCR included in the analysis (*P. anomala* 1, and *P. anomala* 6; Figure 2-3), this GPAT copy was determined to be nearly identical to *P. anomala* 1 in exon and intron sequence. The two sequences differed by only one nucleotide out of more than 1,700 bp, and this difference is likely due to PCR error. Based on the sequence of the intron directly downstream of the exon containing the GAR5 primer (Figure 2-1, 2-2B), *P. anomala* was determined to contain a large insertion that is not present in any of the other *Paeonia* species investigated in this study.

DISCUSSION

For nearly all of the 13 *Paeonia* species, more than one distinct type of GPAT sequence was identified, and, in most cases, within a species the GPAT clones are monophyletic, with a few exceptions (Figure 2-3). In subsection *Foliolatae* of section *Paeonia*, neither the tetraploid *P. mairei* nor *P. obovata* are monophyletic. This is likely due to the history of polyploidy in both species. Furthermore, within section *Moutan*, although each subsection is monophyletic, *P. delavayi*, *P. lutea*, and *P. suffruticosa* ssp.

spontanea are not. Additionally, although each subsection is monophyletic, section *Paeonia* is paraphyletic, with section *Onaepia* resolved as the sister group of subsection *Foliolatae*. This suggests that the history of duplication and deletion of the GPAT gene in *Paeonia* is quite dynamic, and, as a result, some paralogous loci have been maintained between species within subsections, and even between sections *Paeonia* and *Onaepia*.

The GPAT gene phylogeny (Figure 2-3) is well resolved, and the relationships therein are strongly supported. Nearly every node on the gene tree has bootstrap support >50%, and most are supported by bootstrap values >90%, suggesting that there is a strong phylogenetic signal in the GPAT data set. However, the sister relationship of *P. californica* of section *Onaepia* and subsection *Foliolatea* of section *Paeonia* (Figure 2-3) was determined to be significantly incongruent with previous phylogenetic hypotheses.

All previous molecular and morphological investigations in the genus *Paeonia* strongly support the monophyly of both section *Paeonia* and section *Onaepia*. The incongruence between the GPAT gene tree and the *Paeonia* 'species phylogeny' could have arisen through multiple mechanisms (Doyle 1992; Maddison 1997), however, in this case it is likely due to the sampling of paralogous loci among species. If there were a duplication of the GPAT gene before diversification of sections *Paeonia* and *Onaepia*, the sampling of paralogous GPAT loci between species of the two sections would be possible. The strongly supported sister relationship of *P. californica* of section *Onaepia* and subsection *Foliolatea* of section *Paeonia* on the GPAT phylogeny (Figure 2-3), is most likely the result of three independent deletions following an ancient duplication of the GPAT gene prior to diversification of sections *Paeonia* and *Onaepia* (Figure 2-4).

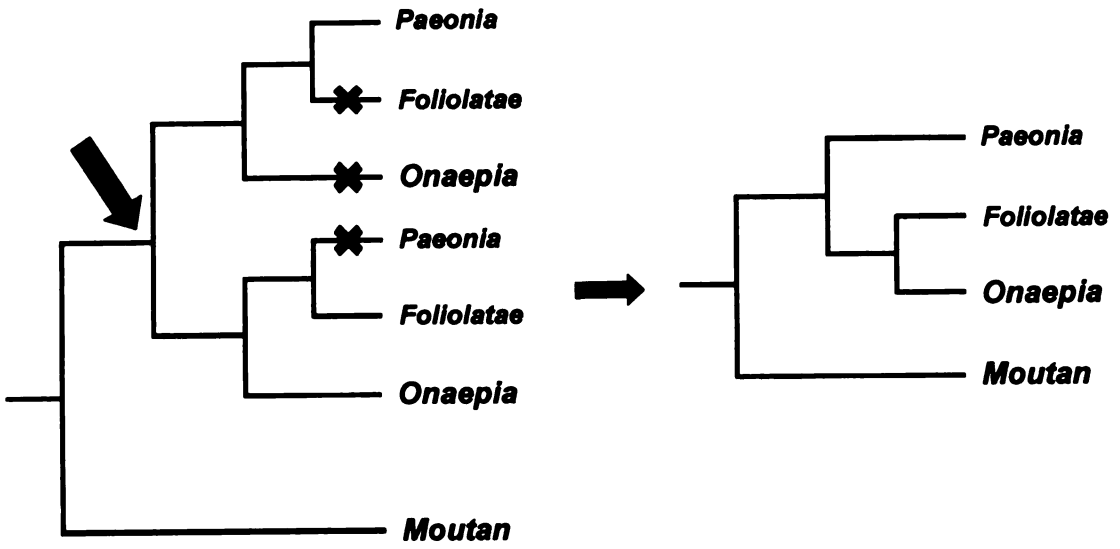


Figure 2-4. Trees depicting the paralogous relationships of the GPAT gene in between section *Paeonia* (subsections *Paeonia* and *Foliolatae*) and section *Onaepia*. The large arrow indicates the gene duplication event, Xs represent independent deletion events, and the small arrow indicates the resulting GPAT gene tree.

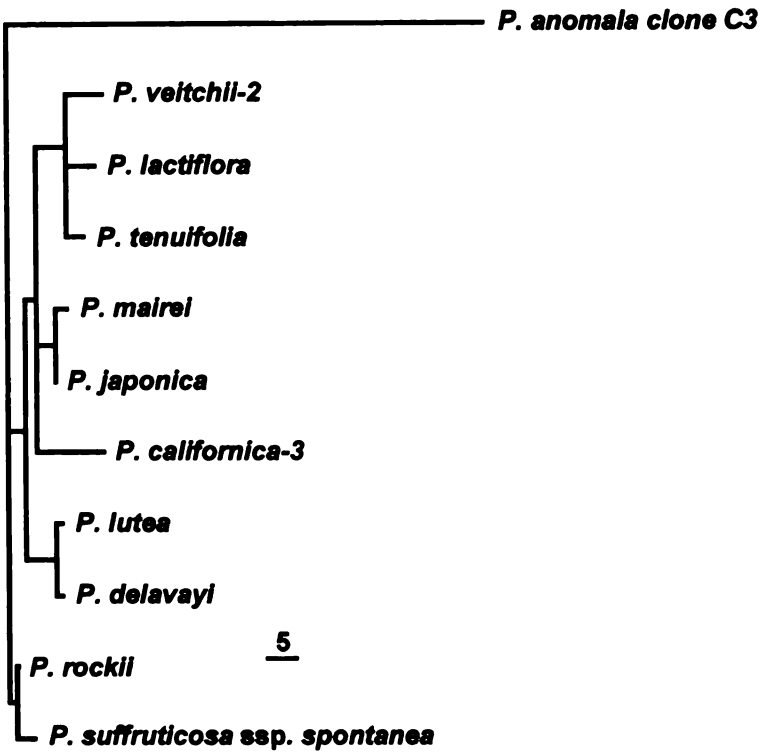


Figure 2-5. GPAT gene phylogeny of 10 *Paeonia* species with the *Paeonia anomala* GPAT pseudogene from genomic clone C3, illustrating an ancient gene duplication event. Strict consensus of four most parsimonious trees. Branch lengths are proportional to the numbers of nucleotide substitutions and are measured by the scale bar.

Based on results from genomic library screening, there is evidence that the GPAT gene indeed has a history of ancient duplication in *P. anomala*. The genomic clone C3 (Figure 2-2A) was determined to be a pseudogene, as evidenced by numerous mutations in both exon and intron regions. Parsimony analysis based on sequence alignment of GPAT sequences from 10 *Paeonia* species with the two recovered pseudogene exons and portions of the flanking introns that could be aligned unambiguously, illustrate the degree of divergence of the GPAT pseudogene (Figure 2-5). Out of a total of ~250 bp of alignable sequence, there are over 81 substitutions along the branch leading to the GPAT pseudogene, indicating that the GPAT gene in *P. anomola* has undergone an ancient duplication, followed by the subsequent silencing of one of the duplicate loci.

As indicated by high identity to sequences available in GenBank, the *Pol* gene, a gene found in retrotransposable elements, was located upstream of one of the GPAT pseudogene exons (Figure 2-2A). Furthermore, after sequencing nearly four kb upstream of the two exons, the resulting nucleotide sequence no longer showed any identity to the GPAT gene. Therefore, it is likely that the insertion of a retrotransposon-like element, interrupting the GPAT gene, was the mechanism responsible for the silencing of the GPAT locus.

Not only is the GPAT gene in *Paeonia* dynamic in its history of duplication and deletion, but there is also plasticity in the intron structure of the gene, as evidenced by the C7 genomic clone (Figure 2-2B). Based on sequences of genomic clone C7, *P. anomala* was found to contain three introns that, in comparison to the structure of the gene in *Arabidopsis*, are dramatically increased in size (Figure 2-2B, 2-1). One of these large introns was identified in all *Paeonia* species (Figure 2-1), and was the primary source of

phylogenetically informative characters in the GPAT phylogeny (Figure 2-3). However, the intron downstream of the 62 bp exon containing the GAR5 PCR primer (Figure 2-2B, 2-1) is also enlarged, and was only identified in *P. anomala*, suggesting that it is the result of a relatively recent insertion event. The third oversized intron, located upstream of the 58 bp exon containing the GAFe5 PCR primer (Figure 2-2B, 2-1), is present in *P. anomala*, and is thought to be present in all *Paeonia* species due to the failure of the GAF1 PCR primer (Figure 2-1) to amplify across this intron.

In conclusion, to effectively use low-copy nuclear genes for low-level phylogeny reconstruction, it is important that mechanisms influencing low-copy nuclear gene evolution are explored. This includes most notably, the influence of duplication and deletion on the topology of gene phylogenies in comparison to the underlying species phylogeny. This study shows a clear example of the reconstruction of an incongruent 'global' phylogeny of *Paeonia*, when compared to the inferred 'species phylogeny' of the genus, due to the sampling of paralogous loci between sections *Paeonia* and *Onaepia*. Therefore, due to paralogy problems between the two sections, the GPAT gene may not be useful for reconstructing the interspecific relationships of the genus as a whole. However, the GPAT gene is a potentially informative independent phylogenetic marker for determining interspecific relationships in *Paeonia* on a more 'local' scale (e.g., within subsection *Paeonia*).

LITERATURE CITED

BHELLA, R. S. and S. L. MACKENZIE. 1994. Nucleotide sequence of a cDNA from *Carthamus tinctorius* encoding a glycerol-3-phosphate acyltransferase. Plant Physiol. **106**:1713-1714.

DOYLE, J. J. and J. L. DOYLE. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem. Bull. **19**: 11-15.

DOYLE, J. J. 1992. Gene trees and species trees – molecular systematics as one-character taxonomy. Syst. Bot. **17**: 144-163.

DOYLE, J. J., V. KANAZIN, and R. C. SHOEMAKER. 1996. Phylogenetic utility of histone H3 intron sequences in the perennial relatives of soybean (*Glycine*: Leguminosae). Mol. Phylog. Evol. **6**: 438-447.

EMSHWILLER, E. and J. J. DOYLE. 1999. Chloroplast-expressed glutamine synthetase (ncpGS): potential utility for phylogenetic studies with an example from *Oxalis* (Oxalidaceae). Mol. Phylog. Evol. **12**: 310-319.

GOTTLIEB, L. D. and V. S. FORD. 1996. Phylogenetic relationships among the sections of *Clarkia* (Onagraceae) inferred from the nucleotide sequences of *PgiC*. Syst. Bot. **21**: 45-62.

ISHIZAKI, O., I. NISHIDA, K. AGATA, G. EGUCHI, and N. MURATA. 1988. Cloning and nucleotide sequence of cDNA for the plastid glycerol-3-phosphate acyltransferase from squash. FEBS Lett. **238**: 424-430.

LI, W.-H. 1997. Molecular Evolution. Sinauer Associates, Sunderland, MA.

MASON-GAMER, R. J., C. F. WEIL, and E. A. KELLOGG. 1998. Granule-bound starch synthase: Structure, function, and phylogenetic utility. Mol. Biol. Evol. **15**: 1658-1673.

MADDISON, W. P. 1997. Gene trees in species trees. Syst. Biol. **46**: 523-536.

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

NISHIDA, I., Y. TASAKA, H. SHIRAISI, and N. MURATA. 1993. The gene and the RNA for the precursor to the plastid-locatedglycerol-3-phosphate acyltransferase of *Arabidopsis thaliana*. Plant Mol. Biol. **21**: 267-277.

PAN, K.-Y. 1979. *Paeonia*. In Flora reipublicae siniccae, vol. 27, 37-59. Science Press, Beijing

PEI, Y.-L. 1993. Studies on the *Paeonia suffruticosa* Andr. Complex. Ph.D. dissertation. Institute of Botany, Chinese Academy of Sciences, Beijing.

SAMBROOK, J., E. F. FRITSCH, and T. MANIATIS. 1989. Molecular cloning: A laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

SANG, T., D. J. CRAWFORD, and T. F. STUESSY. 1995. Documentation of reticulate evolution in peonies (*Paeonia*) using internal transcribed spacer sequences of nuclear ribosomal DNA: Implications for biogeography and concerted evolution. Proc. Natl. Acad. Sci. USA. **92**: 6813-6817.

SANG, T., D. J. CRAWFORD, and T. F. STUESSY. 1997. Chloroplast DNA phylogeny, reticulate evolution, and biogeography of *Paeonia* (Paeoniaceae). Amer. J. Bot. **84**: 1120-1136.

SANG, T., DONOGHUE, M. J., and D. ZHANG. 1997. Evolution of alcohol dehydrogenase genes in peonies (*Paeonia*): phylogenetic relationships of putative nonhybrid species. Mol. Biol. Evol. **14**: 994-1007.

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

SMALL, R. L., J. A. RYBURN, R. C. CRONN, T. SEELANAN, and J. F. WENDEL. 1998. The tortoise and the hare: choosing between noncoding plastome and nuclear *Adh* sequences for phylogeny reconstruction in a recently diverged plant group. Amer. J. Bot. **85**: 1301-1315.

STERN, F. C. 1946. A study of the genus *Paeonia*. Royal Horticultural Society, London.

SWOFFORD, D. L. 1998. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other methods). Version 4. Sinauer Associates, Sunderland, MA.

TEMPLETON, A. R. 1983. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. Evolution **37**: 221-244.

TZANOUDAKIS, D. 1983. Karyotypes of four wild *Paeonia* species from Greece. Nordic J. Bot. **3**: 307-318.

WANG, X.-Q., D. C. TANK, T. SANG. 2000. Phylogeny and divergence times in Pinaceae: Evidence from three genomes. Mol. Biol. Evol. **17**: 773-781.

WEBER, S., F. P. WOLTER, F. BUCK, M. FRENTZEN and E. HEINZ. 1991. Purification and cDNA sequencing of an oleate-selective acyl-ACO : sn-glycerol-3-phosphate acyltransferase from pea chloroplasts. Plant Mol. Biol. **17**: 1067-1076.

APPENDIX

PHYLOGENY AND DIVERGENCE TIMES IN PINACEAE: EVIDENCE FROM THREE GENOMES

Phylogeny and Divergence Times in Pinaceae: Evidence from Three Genomes

Xiao-Quan Wang, David C. Tank,† and Tao Sang†*

*Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing, China; and †Department of Botany and Plant Pathology, Michigan State University

In Pinaceae, the chloroplast, mitochondrial, and nuclear genomes are paternally, maternally, and biparentally inherited, respectively. Examining congruence and incongruence of gene phylogenies among the three genomes should provide insights into phylogenetic relationships within the family. Here we studied intergeneric relationships of Pinaceae using sequences of the chloroplast *matK* gene, the mitochondrial *nad5* gene, and the low-copy nuclear gene 4CL. The 4CL gene may exist as a single copy in some species of Pinaceae, but it constitutes a small gene family with two or three members in others. Duplication and deletion of the 4CL gene occurred at a tempo such that paralogous loci are maintained within but not between genera. Exons of the 4CL gene have diverged approximately twice as fast as the *matK* gene and five times more rapidly than the *nad5* gene. The partition-homogeneity test indicates that the three data sets are homogeneous. A combined analysis of the three gene sequences generated a well-resolved and strongly supported phylogeny. The combined phylogeny, which is topologically congruent with the three individual gene trees based on the Templeton test, is likely to represent the organismal phylogeny of Pinaceae. This phylogeny agrees to a certain extent with previous phylogenetic hypotheses based on morphological, anatomical, and immunological data. Disagreement between the previous hypotheses and the three-genome phylogeny suggests that morphology of both vegetative and reproductive organs has undergone convergent evolution within the pine family. The strongly supported monophyly of *Nothotsuga longibracteata*, *Tsuga mertensiana*, and *Tsuga canadensis* on all three gene phylogenies provides evidence against previous hypotheses of intergeneric hybrid origins of *N. longibracteata* and *T. mertensiana*. Divergence times of the genera were estimated based on sequence divergence of the *matK* gene, and they correspond well with the fossil record.

Introduction

A plant cell has one nuclear and two organellar (chloroplast and mitochondrial) genomes. Genes from the different genomes may have distinct phylogenies as a result of different inheritance pathways and differential responses to processes such as lineage sorting, gene duplication/deletion, lateral gene transfer, and hybrid speciation (Doyle 1997; Maddison 1997; Wendel and Doyle 1998). Conversely, congruent phylogenies among the three genomes could suggest strongly that the gene trees are also congruent with the single underlying phylogeny—the species phylogeny. Therefore, comparison of gene phylogenies of the three genomes will provide an opportunity for robust reconstructions of complex plant phylogenies (e.g., Qiu and Palmer 1999).

Inheritance pathways of the three genomes in the pine family (Pinaceae) are strikingly different; the chloroplast, mitochondrial, and nuclear genomes are paternally, maternally, and biparentally inherited, respectively (Gillham 1994; Hipkins, Krutovskii, and Strauss 1994; Mogensen 1996). Pinaceae, comprising 11 genera and more than 200 species (Farjon 1998), is the largest extant family of gymnosperms. Many species of the pine family constitute the major forest elements in the northern temperate region. Due to morphological convergence within the family, Pinaceae has been a phylogenetically complex group (Hart 1987; Farjon 1990). Phylogenetic relationships of two monotypic genera, *Cathaya* and *Nothotsuga*, and *Tsuga mertensiana* (once recognized as a monotypic genus, *Hesperopeuce*), are particularly controversial because each of them shares morphological features with several other genera (Frankis 1988; Page 1988; Lin, Hu, and Wang 1995; Wang, Han, and Hong 1998*a*). *Nothotsuga longibracteata* and *T. mertensiana* were further hypothesized to be intergeneric hybrids based on their morphological intermediacy (Van Campo-Duplan and Gaussen 1948; Gaussen 1966).

Previous molecular phylogenetic studies of intergeneric relationships in Pinaceae were based primarily on the chloroplast genome. The phylogeny generated from *rbcL* gene sequences was poorly resolved and contradicts the phylogeny generated from PCR restriction fragment length polymorphisms of six chloroplast genes (Tsumura et al. 1995; Wang, Han, and Hong 1998*b*). Each previous molecular phylogenetic study involving Pinaceae based on nuclear ribosomal genes sampled only three or four genera (Chaw et al. 1997; Stefanovic et al. 1998; Gernandt and Liston 1999).

In this study, we included all the extant genera of Pinaceae and compared gene phylogenies of the three genomes in order to clarify intergeneric relationships. We chose a rapidly evolving gene, *matK*, for reconstruction of the phylogeny of the chloroplast genome (Johnson and Soltis 1994, 1995; Steele and Vilgalys 1994). For the mitochondrial genome, which has slow rates of nucleotide substitutions (Hiesel, Haeseler, and Brennicke 1994; Laroche et al. 1997), we sequenced an intron of the *nad5* gene encoding subunit 5 of NADH dehydrogenase. The nuclear genome of conifers is large in size and complex in organization, and genes usually exist in large gene families (Perry and Furnier 1996; Kinlaw and Neale 1997; Murray 1998). The questions of how dynamically gene duplication/deletion occurs in co-

Key words: Pinaceae, chloroplast *matK*, mitochondrial *nad5*, nuclear gene 4CL, gene duplication and deletion, molecular clock.

Address for correspondence and reprints: Tao Sang, Department of Botany and Plant Pathology, Michigan State University, East Lansing, Michigan 48824. E-mail: sang@pilot.msu.edu.

Mol. Biol. Evol. 17(5):773–781. 2000
 ISSN: 0737-4038

Table 1
Collection localities of Species of Pinaceae and *Cycas* Sampled for DNA Sequencing in this Study

Species	Collection Locality
Abies beshanzuensis Wu	Longquan, Zhejiang, China
Abies firma Sieb. et Zucc.	Botanic Garden, Institute of Botany, Beijing
Abies holoyphylla Maxim.	Botanic Garden, Institute of Botany, Beijing
Cathaya argyrophylla Chun et Kuang	Huaping, Guangxi, China
Cedrus atlantica Manetti	Michigan State University, East Lansing, Mich.
Keteleeria evelyniana Mast.	Botanic Garden, Institute of Botany, Kunming
Larix gmelini (Rupr.) Rupr.	Botanic Garden, Institute of Botany, Beijing
Nothotsuga longibracteata Hu ex Page	Xinning, Hunan, China
Picea smithiana (Wall.) Boiss.	Botanic Garden, Institute of Botany, Beijing
Pinus armandi Franch.	Botanic Garden, Institute of Botany, Beijing
Pinus banksiana Lamb.	Botanic Garden, Institute of Botany, Beijing
Psuedolarix amabilis (Nelson) Rehd.	Botanic Garden, Institute of Botany, Kunming
Pseudotsuga menziesii (Mirbel) Franco	Botanic Garden, Institute of Botany, Beijing
Pseudotsuga sinensis Dode	Botanic Garden, Institute of Botany, Kunming
Tsuga canadensis Carr.	Michigan State University, East Lansing, Mich.
Tsuga mertensiana (Bong.) Rydb.	Mt. Hood, Oreg.
Cycas panzhihuaensis Zhou et Yang	Panzhihua, Sichuan, China

nifers and how this will affect phylogenetic utility of nuclear genes remain open (Kinlaw and Neale 1997). Single- or low-copy nuclear genes have been increasingly used for phylogenetic studies on angiosperms (e.g., Doyle, Kanazin, and Shoemaker 1996; Gottlieb and Ford 1996; Sang, Donoghue, and Zhang 1997; Mason-Gamer, Weil, and Kellogg 1998; Small et al. 1998). In the present study, we chose the low-copy nuclear gene 4CL, encoding 4-coumarate : coenzyme A ligase in the lignin biosynthetic pathway (Zhang and Chiang 1997), for study of gene duplication/deletion and inference of the phylogeny of Pinaceae from the nuclear genome.

In addition to reconstructing phylogenetic relationships, we estimated divergence times among the genera of Pinaceae using a molecular clock. It has been shown for both plants and animals that divergence times calculated from the molecular clock may not be concordant with those based on the fossil record (e.g., Martin, Gierl, and Saedler 1989; Wolfe et al. 1989; Bromham et al. 1998). The abundant fossil record of the pine family (Florin 1963) offers an excellent opportunity for the comparison of these two approaches to determine divergence times.

Materials and Methods

All 11 recognized genera of Pinaceae were sampled, including *Abies* (fir), *Cathaya, Cedrus* (cedar), *Keteleeria, Larix* (larch), *Nothotsuga, Picea* (spruce), *Pinus* (pine), *Pseudolarix* (golden larch), *Pseudotsuga* (Douglas-fir), and *Tsuga* (hemlock). Sampling localities are given in table 1. Voucher specimens have been deposited in the herbaria of the Institute of Botany, Beijing, and Michigan State University. Total DNA was isolated from fresh leaves using the CTAB method (Doyle and Doyle 1987) and purified with a Wizard DNA Clean-up System (Promega).

Genes were amplified through the following PCR cycles: (1) 70°C, 4 min; (2–5) 94°C, 1 min; 48–55°C, 30 s; 72°C, 2 min; (6–36) 94°C, 20 s; 48–55°C, 30 s; 72°C, 2 min; and (37) 72°C, 5 min. Primers for amplifying the *matK* gene are *trnK*-3914F and *trnK*-2R (Johnson and Soltis 1995) with an additional forward primer, *trnK*F1 (5′-TACTGATCAGAAGTTAAGAGC). For the *nad5* gene, the forward primer *nad5*-aF (5′-GGAAATGTTTGATGCTTCTTGGG) and the reverse primer *nad5*-bR (5′-CTGATCCAAAATCACCTACTCG) are located on exons a and b, respectively. For the 4CL gene, the forward primers 4CLpF2 (5′-AGAGTVGCGGAATTCGCAG) and 4CLpF3 (5′-CCAATCCTTTYTACAAGCCG) are located on exon 1, and the reverse primers 4CLpR2 (5′-TTTGAGCGTTMCGGACGAC) and 4CLpR3 (5′-CGGGGAARGGCTYCTTTGC) are located on exon 2.

PCR products of *matK* and *nad5* genes were purified using Genclean (Bio 101). PCR products of the nuclear 4CL gene were cloned with a TA cloning kit (Invitrogen). For each species, 10–30 clones were screened by examining restriction site or sequence (from one primer) variation (Sang, Donoghue, and Zhang 1997). Distinct clones were fully sequenced and included in the phylogenetic analyses. Sequencing was done on an ABI 373 automated DNA sequencer using the Dye Terminator Cycle Sequencing reaction kit (PE Applied Biosystems). Sequences have been deposited in GenBank under accession numbers AF143412–AF143425 (*nad5*), AF143427–AF143441 (*matK*), and AF144499–AF144529 (4CL). Additional sequences obtained from GenBank include *matK* genes of *Pinus thunbergii* (D11467) (Tsudzuki et al. 1992), *Pinus contorta* (X57097) (Lidholm and Gustafsson 1991), *Picea glauca* (AF059341), *Picea rubens* (AF059342), and *Picea mariana* (AF059343) and 4CL genes of *Pinus taeda* (U39404 and U39405) (Zhang and Chiang 1997) and *Arabidopsis thaliana* (U18675) (Lee et al. 1995).

Sequence alignments were made with CLUSTAL W (Thompson, Higgins, and Gibson 1994) and refined manually. A few regions in the 4CL intron could not be aligned unambiguously and were excluded from the analyses. Parsimony, as implemented in PAUP*, version 4.0 (Swofford 1998), was used to infer phylogenies based on nucleotide substitutions in aligned sequences.

Unweighted parsimony analyses were performed by heuristic search with tree bisection-reconnection (TBR) branch swapping, the MULPARS option, ACCTRAN optimization, and 1,000 random-addition replicates for the 4CL data set, or by branch-and-bound search with the options of Multree and farthest sequence addition for the *matK*, *nad5*, and combined data sets. Bootstrap analyses (Felsenstein 1985) were carried out with 1,000 replications of heuristic search with simple taxon addition, while all trees were saved. *Cycas* was chosen as the outgroup for phylogenetic analyses of the *matK* and *nad5* sequences because sequence divergence of the *rbcL* gene is lower between Pinaceae and *Cycas* than between Pinaceae and Podocarpaceae or Araucariaceae (Wang, Han, and Hong 1998*a*). However, we were unable to amplify the 4CL gene from *Cycas*. *Arabidopsis* was used as the outgroup when only exon sequences of the 4CL gene were analyzed. In the resulting parsimony tree, *Cedrus* formed the sister group to the remaining genera with 78% bootstrap support. The same basal relationship of *Cedrus* was obtained from both *matK* and *nad5* phylogenies when *Cycas* was used as the outgroup (see *Results*). Thus, *Cedrus* was chosen as the functional outgroup for further parsimony analysis of both exon and intron regions of the 4CL gene.

Congruence among the three data sets was examined with the partition-homogeneity test (Farris et al. 1995), implemented in PAUP*, version 4.0. For the purposes of this test, data sets of the three genes were reduced so that they shared the same set of 13 taxa. In the reduced data sets, each genus was represented by a single species, except for *Pinus* and *Tsuga*, of which subgenera were also represented. In the reduction of the 4CL data set, a single clone was chosen randomly to represent a species with multiple distinct 4CL sequences. *Cedrus* was used as the functional outgroup, while *Cycas* was excluded from the *matK* and *nad5* data sets to maintain consistency with the 4CL data set. The tests were performed with 100 replications of heuristic search with TBR branch swapping. Topological congruence between the gene trees was evaluated with the Templeton (1983) test, implemented in PAUP*, version 4.0.

Maximum-likelihood analyses were performed using PAUP*, version 4.0. The program Modeltest, version 2.1 (Posada and Crandall 1998), was utilized to find the model of sequence evolution that best fit each data set by the hierarchical likelihood ratio (LR) test (α = 0.05). When the models of sequence evolution are nested, the LR test statistic is distributed as χ^2 with degrees of freedom equal to the number of free parameters between the two models (Goldman 1993). Once the best sequence evolution model was determined (table 2), maximum-likelihood tree searches were performed for each data set. The molecular-clock hypothesis was tested with the LR test by calculating the log likelihood score of the chosen model with the molecular clock enforced and comparing it with the log likelihood score without the molecular clock enforced (Muse and Weir 1992; Baldwin and Sanderson 1998). The number of degrees of freedom is equivalent to the number of terminals minus two (Sorhannus and Van Bell 1999).

Table 2
Sequence Evolution Models Best Fit to Each Data Set as Determined by Hierarchical Likelihood Ratio Tests

Data Set	Models
matK	K3Puf+Γ
nad5	K3P+I+Γ
4CL	K2P+Γ
Three-gene combined	K3Puf+Γ

NOTE.—The models of DNA substitution are the Kimura (1986) two-parameter model (K2P); the Kimura (1981) three-parameter model (K3P), K3P with unequal base frequencies (K3Puf, Kimura 1981). Λ = shape parameter of the gamma distribution (estimated via maximum-likelihood); I = proportion of invariable sites (estimated via maximum-likelihood).

Results

The aligned *matK* sequences were 1,551 bp in length, of which 545 nucleotide sites were variable and 210 were parsimony-informative. Parsimony analysis generated a single most-parsimonious tree with a tree length of 778, a consistency index (CI) of 0.80, and a retention index (RI) of 0.76 (fig. 1*A*). The aligned sequences of the *nad5* gene included 285 bp of exon and 1,042 bp of intron, of which 141 nucleotide sites were variable and 54 were parsimony-informative. Parsimony analysis yielded three equally most parsimonious trees (tree length = 184, CI = 0.80, RI = 0.70). The parsimonious tree that is topologically identical to the maximum-likelihood (ML) tree is shown in figure 1*B*. Although the basal position of *Cedrus* collapsed on the strict consensus of the three parsimonious trees, *Cedrus* is the sister group of the remaining genera of the family on the *nad5* ML tree. This result supports the utility of *Cedrus* as a functional outgroup in analyses of the 4CL and combined data sets.

After screening 10–30 4CL clones for each species, only one type of clone was found for *Cathaya argyrophylla*, *Cedrus atlantica*, *T. mertensiana*, *Keteleeria evelyniana*, *Picea smithiana*, and *N. longibracteata*. Two types of clones were identified for each of the following species: *Abies holophylla*, *Larix gmelini*, *Pinus banksiana*, *Pseudolarix amabilis*, and *Tsuga canadensis*. Three types of clones were isolated for each of the following species: *Abies firma*, *Abies beshanzuensis*, *Pinus armandi*, *Pseudotsuga menziesii*, and *Pseudotsuga sinensis*. The 4CL data set contained 827 bp of exon and 126 bp of alignable intron sequences, of which 360 sites were variable and 264 were parsimony-informative. Parsimony analysis resulted in six equally most parsimonious trees (tree length = 649, CI = 0.71, RI = 0.86). The parsimonious tree that is topologically identical to the ML tree is shown in figure 1*C*.

When the three data sets were reduced to 13 taxa for the homogeneity tests, the *matK*, *nad5*, and 4CL data sets contained 131, 47, and 153 parsimony-informative sites, respectively. The partition-homogeneity tests indicated that all pairs of the three data sets are congruent (P = 0.17 for *matK*-*nad5*; P = 0.20 for *matK*-4CL; P = 0.17 for *nad5*-4CL). Therefore, the three data sets were combined for further phylogenetic analysis. Three equally most parsimonious trees (tree length = 1,110,

FIG. 1.—Phylogenies of chloroplast *matK*, mitochondrial *nad5*, and nuclear 4CL genes of Pinaceae. *A*, *matK* gene phylogeny. The single most parsimonious tree (tree length = 778, consistency index [CI] = 0.80, retention index [RI] = 0.76). *B*, *nad5* gene phylogeny. One of three equally most parsimonious trees (tree length = 184, CI = 0.80, RI = 0.70). *C*, 4CL gene phylogeny. One of six equally most parsimonious trees (tree length = 649, CI = 0.71, RI = 0.86). Small numbers following a species name indicate clone numbers. Numbers associated with branches are bootstrap percentages greater than 50%. When multiple parsimonious trees are found from the *nad5* and 4CL data sets, the one with the same topology as the maximum-likelihood tree is shown. * = branch collapses on the strict consensus. Branch lengths are proportional to the numbers of nucleotide substitutions and are measured by scale bars.

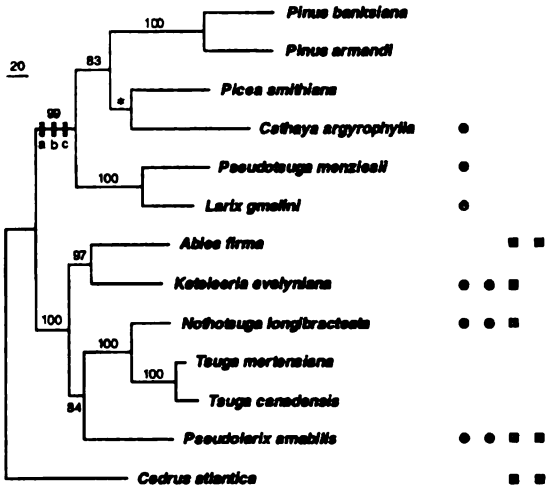


FIG. 2.—Phylogeny of Pinaceae based on combined sequences of three genes, *matK*, *nad5*, and 4CL; one of three equally most parsimonious trees (tree length = 1,110, CI = 0.78, RI = 0.68) which is topologically identical to the maximum-likelihood tree. * = branch collapses on the strict consensus. Numbers associated with branches are bootstrap percentages greater than 50%. Branch lengths are proportional to the numbers of nucleotide substitutions and are measured by the scale bar. Synapomorphies supporting a branch are indicated by black bars: (*a*) absence of resin vesicles in seed coat, (*b*) absence of narrowed, pedicellate base of seed scales, and (*c*) presence of two resin canals in vascular cylinder of young taproot. Four morphological characters that may have undergone parallel changes are labeled next to the species names: gray circle, cones on leaved peduncles; black circle, male strobili in clusters from a single bud; gray square, erect position of mature cones; black square, seed scale abscission.

CI = 0.78, RI = 0.68) were obtained from the combined data set. The parsimonious tree that is topologically identical to the ML tree is shown in figure 2. For each gene, the average sequence divergence among these 13 taxa was estimated with the Jukes-Cantor model (Jukes and Cantor 1969) as 10.93% for the 4CL exons, 5.62% for the *matK* gene, and 2.21% and 2.46% for the exon and intron of the *nad5* gene, respectively.

The *matK* phylogeny is topologically congruent with the tree resulting from the combined analysis (figs. 1*A* and 2). Topological incongruence between the *nad5* and the combined tree, which is supported by bootstrap values higher than 50% on both trees, involves only the position of *Pseudolarix* (figs. 1*B* and 2). The Templeton test was performed on the *nad5* data set, while the topology of the combined tree was used as a constraint. The analysis with the constraint did not lead to an increase in tree length, and thus the topological incongruence was not significant. While topological incongruence between the 4CL and the combined trees involves the positions of *Cathaya, Keteleeria,* and *Pseudolarix* (figs. 1*C* and 2), bootstrap support is found only for *Cathaya* on the 4CL tree. Using the topology of the combined tree as a constraint (fig. 2), the Templeton test indicated that the incongruence was not significant (Ts = 35.0, N = 9, P = 0.10).

Results of the LR test of the molecular-clock hypothesis for the reduced data sets (each containing 13 taxa) of the three genes are as follows: *matK*, -2 ln LR = 24.64, df = 11, $0.025 > P > 0.01$; *nad5*, -2 ln LR = 22.56, df = 11, $0.025 > P > 0.01$; and 4CL, -2 ln LR = 39.50, df = 11, $P < 0.001$. Because the molecular clock of the *matK* and *nad5* genes cannot be rejected at the significance level of $P = 0.01$, sequence divergence of these two genes may be useful in estimating divergence times. However, when the molecular clock was enforced, ML analyses of the *matK* and *nad5* data sets yielded trees (not shown) with topologies different from the parsimonious trees. On the *matK* ML tree with molecular clock (the ML-MC tree), *Cedrus* formed a sister group with the clade containing *Abies, Keteleeria, Nothotsuga, Tsuga,* and *Pseudolarix*. On the *nad5* ML-MC tree, *Cedrus* formed a sister group with the clade containing *Pinus, Picea, Cathaya, Pseudotsuga,* and *Larix,* and the clade of *Larix* and *Pseudotsuga* became the sister group of the remaining genera of the family.

By excluding *Cedrus* from the *matK* data set and rooting the tree between the next two major basal clades of the three-gene phylogeny (fig. 2), the resulting ML-MC tree (fig. 3) has the same topology as the *matK* (fig. 1*A*) and three-gene phylogenies. The molecular clock for the remaining sequences could not be rejected at $P = 0.025$ (-2 ln LR = 19.78, df = 10, $0.05 > P > 0.025$). Because *Cedrus* has the shortest branch length on the *matK* gene tree (fig. 1*A*), the slow divergence rate of the *matK* sequences of *Cedrus* may have contributed in part to the rate heterogeneity of the *matK* data set. For the *nad5* data set, although the molecular clock could not be rejected after excluding *Cedrus*, the clade of *Larix* and *Pseudotsuga* remained as the sister group of the rest of the genera (tree not shown). Therefore, only *matK* sequences were used to estimate divergence times of all genera except *Cedrus*. Branch lengths were estimated by ML with the molecular clock enforced (fig. 3).

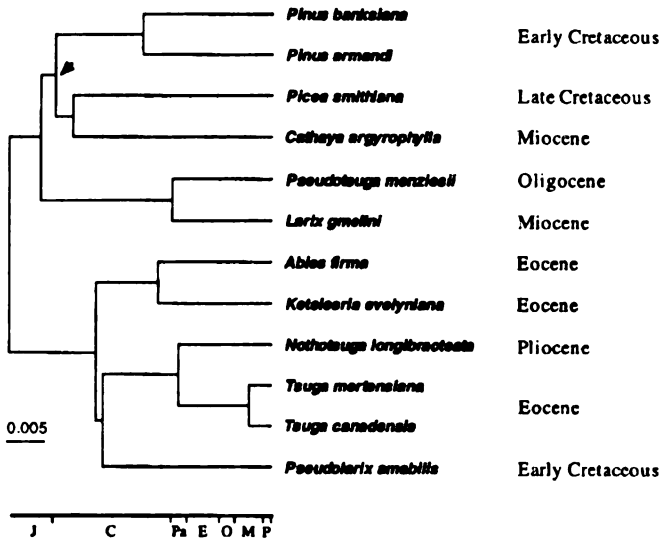


FIG. 3.—Maximum-likelihood tree of Pinaceae based on *matK* sequences with a molecular clock enforced. The earliest fossil record of each genus (not necessarily the species sampled in this study) is indicated. Branch lengths are proportional to sequence divergence estimated by maximum-likelihood and are measured by the scale bar. The geological timescale was calculated from the branch lengths according to the molecular clock: J, Jurassic; C, Cretaceous; Pa, Paleocene; E, Eocene; O, Oligocene; M, Miocene; P, Pliocene. The arrow indicates the point at which the molecular clock is calibrated (140 MYA).

Discussion

Evolution of 4CL Gene

A better understanding of the dynamics of gene duplication/deletion will provide insights into evolution of the nuclear genome, as well as the phylogenetic utility of low-copy nuclear genes (Morton, Gaut, and Clegg 1996; Clegg, Cummings, and Durbin 1997). Two 4CL loci were previously found in *P. taeda* (Zhang and Chiang 1997), and their sequences formed a monophyletic group on the 4CL phylogeny (fig. 1*C*). This study identified as many as three distinct clones from individuals of some species. Observed sequence divergence between clones isolated from the same individual ranged from 3 bp (between *P. armandi* 1 and *P. armandi* 11) to 57 bp (between *P. sinensis* 9 and *P. sinensis* 17). Distinct clones isolated from an individual plant may represent different loci or allelic variation. Given that the two 4CL loci isolated previously from *P. taeda* differ by only 2 bp in the partial sequence analyzed here, different sequences cloned from a species in this study, which differ by at least 3 bp, could also represent different 4CL loci. Although only one type of 4CL sequence has been found in a number of species, it is still possible that additional loci in some of the species remain unidentified due to PCR selection (Wagner et al. 1994). Apparently, the 4CL gene of Pinaceae may exist as a single copy in some species, but it constitutes a small gene family with two or three members in others.

It is remarkable that all 4CL clones from each genus form a strongly supported monophyletic group (fig. 1*C*). Sequences cloned from a species, however, do not necessarily form a monophyletic group. Notably, two or three types of sequences were cloned from each of the three *Abies* species. They group into two strongly supported clades, which may represent a gene duplication prior to the diversification of the three *Abies* species. A similar pattern was also found for the two *Pseudotsuga* species. These results indicate that the 4CL gene has a tempo of duplication/deletion cycles such that paralogous loci are maintained between species but not between genera. Therefore, this gene can serve as an efficient phylogenetic marker for studying relationships at or above the intergeneric level. However, caution must be exercised in distinguishing paralogy and orthology when the 4CL gene is used for phylogenetic studies at the interspecific level.

Of the three genes, the nuclear 4CL gene evolved most rapidly. The average sequence divergence of the exon region of the 4CL gene is approximately twice as high as that of the chloroplast *matK* gene and five times as high as that of the mitochondrial *nad5* gene. Because the nuclear ribosomal DNA internal transcribed spacers exhibit a high level of length variation and exist in multiple diverged copies in Pinaceae (Liston et al. 1996; Gernandt and Liston 1999), low-copy nuclear genes will provide useful alternative markers for phylogenetic reconstructions at the intergeneric and interspecific levels. The *matK* gene diverged about twice as fast as the *rbcL* gene (Wang, Han, and Hong 1998*a*) in Pinaceae, which is similar to the rate differences between these two chloroplast genes in angiosperms (Johnson and Soltis 1994, 1995; Steele and Vilgalys 1994). The higher evolutionary rate of the *matK* gene may be responsible for the better resolution and support in the *matK* phylogeny

than in the *rbcL* phylogeny (Wang, Han, and Hong 1998*a*) of Pinaceae. Sequences of the mitochondrial *nad5* gene, including a large intron, have diverged most slowly, concordant with previous observations of relative sequence divergence rates among the three plant genomes (Palmer 1992). Nevertheless, the *nad5* gene tree has offered reasonable resolution among genera of Pinaceae (fig. 1*B*).

Phylogeny of Pinaceae and Evolution of Morphological Characters

Despite the different inheritance pathways of the three genomes, the *matK*, *nad5*, and 4CL data sets are congruent based on the homogeneity test. Although the three gene trees are not identical in topology, incongruence among them is not supported by high bootstrap values, and the three gene trees are topologically congruent with the combined tree based on the Templeton test. Therefore, the tree resulting from the combined analysis is very likely to represent the true intergeneric relationships of Pinaceae. Furthermore, the strongly supported monophyly of *N. longibracteata*, *T. mertensiana*, and *T. canadensis* on all three gene phylogenies provides evidence against previous hypotheses of the hybrid origins of *N. longibracteata* (between *Tsuga* and *Keteleeria*) and *T. mertensiana* (between *Tsuga* and *Picea*) (Van Campo-Duplan and Gaussen 1948; Gaussen 1966). If these were intergeneric hybrids, they would likely have significantly incongruent positions between the chloroplast (paternal) and mitochondrial (maternal) gene phylogenies (Sang, Crawford, and Stuessy 1997; Wendel and Doyle 1998).

The three-genome phylogeny is similar to the phenogram generated from immunological data (Price, Olsen-Stojkovich, and Lowenstein 1987) except for the position of *Cedrus*. The immunological phenogram, which did not sample *Cathaya* or *Nothotsuga*, placed *Cedrus* and *Abies* as sister genera. In contrast, a sister group relationship between *Cedrus* and the rest of the family is revealed here by the *matK* phylogeny (with 84% bootstrap support), the 4CL exon sequences (with 78% bootstrap support with *Arabidopsis* as the outgroup), and the ML tree of the *nad5* gene.

In comparison with the commonly accepted classification systems of Pinaceae, the three-genome phylogeny largely agrees with the classification based on the number and position of resin canals in the central vascular cylinder of the young taproot, which divided the family into two major groups: Cédrées, containing *Abies*, *Cedrus*, *Keteleeria*, *Pseudolarix*, and *Tsuga*, and Pinées, containing *Larix*, *Picea*, *Pinus*, and *Pseudotsuga* (Van Tieghem 1891). The three-genome phylogeny, however, differs markedly from the conventional classification of Pinaceae, which recognizes three subfamilies: Pinoideae (*Pinus*), Laricoideae (*Cedrus*, *Larix*, and *Pseudolarix*), and Abietoideae, consisting of the remaining genera (Melchior and Werdermann 1954). Our results support the previous speculation that shoot and foliage morphology, on which the classification is based, has undergone considerable convergent evolution within Pinaceae (Frankis 1988).

The three-genome phylogeny agrees to a certain extent with the phylogenetic hypothesis based on combined evidence from morphology of both vegetative and reproductive organs, wood and root anatomy, and immunological data (Farjon 1990). By labeling the characters that Farjon (1990) used to define major groups on the three-genome phylogeny, both synapomorphies and parallelisms are illustrated (fig. 2). Synapomorphies, which support the clade of *Cathaya*, *Larix*, *Picea*, *Pinus*, and *Pseudotsuga*, include absence of resin vesicles in the seed coat; absence of a narrowed, pedicellate base of seed scales; and presence of two resin canals in the vascular cylinder of the young taproot. In contrast, assuming homology of the morphological feature "cones on leaved peduncles" leads to grouping *Cathaya* with *Larix* and *Pseudotsuga*. This character, together with "male strobili in clusters from a single bud," grouped *Keteleeria*, *Pseudolarix*, and *Nothotsuga* together. Two characters, "seed scale abscission" and "erect position of mature cones," were mainly responsible for grouping *Abies* and *Cedrus*. These results suggest that the morphology of reproductive organs may have also undergone convergent evolution.

Divergence Times

When the molecular clock was used to estimate divergence times in Pinaceae, *Cedrus* was excluded because its *matK* gene appeared to have diverged at a slower rate. Even when *Cedrus* was excluded, there still existed a certain degree of rate heterogeneity among the remaining *matK* sequences ($0.05 > P > 0.025$). When *Keteleeria*, which has the second shortest branch on the *matK* gene tree (fig. 1*A*), was also excluded, the LR test could not reject the molecular clock ($-2 \ln \text{LR} = 15.69$, $df = 9$, $P > 0.05$). Exclusion of *Keteleeria*, however, had little impact on the estimated divergence times of the remaining genera and did not alter the estimated divergence times at the broad geological timescale where the molecular clock and fossil record were compared. Therefore, the data set that still contains *Keteleeria* was used for estimating divergence times and for further comparison with the fossil record.

Pinaceae has one of the most extensive fossil records of extant plant families. Among genera of Pinaceae, *Pinus* has the the best fossil record, dating back to the early Cretaceous (Miller 1977; Florin 1963). Thus, we calibrated the molecular clock by using 140 MYA as the time when *Pinus* diverged from the other genera (Savard et al. 1994). The geological timescale is estimated accordingly along the branch length of the tree (fig. 3). The earliest fossil records for the genera (Miller 1977, 1998; Florin 1963; Farjon 1990; LePage and Basinger 1995*a*, 1995*b*) are labeled on the *matK* ML-MC tree (fig. 3).

Remarkably, divergence times estimated from the molecular clock correspond well with the fossil record for the majority of the genera. Of four genera, *Pinus*, *Picea*, *Cathaya*, and *Pseudolarix*, which became estab-

lished in the early and middle Cretaceous according to the ML-MC tree, three have fossil records from the Cretaceous. Only *Cathaya*, currently endemic to China, has a much more recent fossil record, first documented in the Miocene. Although the divergence time of *Cedrus* could not be estimated directly, its basal position in Pinaceae revealed by the three genes is concordant with its fossil record from the early Cretaceous (Arnold 1953). The ML-MC tree suggests that the next period of major diversification within the pine family was around the Paleocene. This corresponds well with the earliest fossil records of *Abies*, *Keteleeria*, *Larix*, and *Tsuga* from the Eocene and that of *Pseudotsuga* from the Oligocene. *Nothotsuga*, however, has a rather recent fossil record, dating back only to the Pliocene. The lack of early fossil records of the monotypic genera *Nothotsuga* and *Cathaya* may be due to their limited historical distributions and/or less extensive studies of fossils at these sites.

Acknowledgments

We thank Zhongchun Luo and Sherry Spencer for providing some of the plant material used in this study; Fang Wang for lab assistance; Xuanli Yao for helpful discussion on the fossil record; and Diane Ferguson, Pam Soltis, and two anonymous reviewers for valuable comments on the manuscript. This study was supported by Michigan State University, the National Natural Science Foundation of China (grant 39391500), and the Chinese Academy of Sciences.

LITERATURE CITED

ARNOLD, C. A. 1953. Silicified plant remains from the Mesozoic and Tertiary of North America. II. Some fossils from northern Alaska. Mich. Acad. Sci. Lett. **38**:9–12.

BALDWIN, B. G., and M. J. SANDERSON. 1998. Age and rate of diversification of the Hawaiian silversword alliance (Compositae) Proc. Natl. Acad. Sci. USA **95**:9402–9406.

BROMHAM, L., A. RAMBAUT, R. FORTEY, A. COOPER, and D. PENNY. 1998. Testing the Cambrian explosion hypothesis by using a molecular dating technique. Proc. Natl. Acad. Sci. USA **95**:12386–12389.

CHAW, S.-M., A. ZHARKIKH, H.-M. SUNG, T.-C. LAU, and W.-H. LI. 1997. Molecular phylogeny of extant gymnosperms and seed plant evolution: analysis of nuclear 18S rRNA sequences. Mol. Biol. Evol. **14**:56–68.

CLEGG, M. T., M. P. CUMMINGS, and M. L. DURBIN. 1997. The evolution of plant nuclear genes. Proc. Natl. Acad. Sci. USA **94**:7791–7798.

DOYLE, J. J. 1997. Trees within trees: genes and species, molecules and morphology. Syst. Biol. **46**:537–553.

DOYLE, J. J., and J. L. DOYLE. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem. Bull. **19**:11–15.

DOYLE, J. J., V. KANAZIN, and R. C. SHOEMAKER. 1996. Phylogenetic utility of histone H3 intron sequences in the perennial relatives of soybean (*Glycine*: Leguminosae). Mol. Phylogenet. Evol. **6**:438–447.

FARJON, A. 1990. Pinaceae. Koleltz Scientific Books, Konigstein, Germany.

———. 1998. World checklist and bibliography of conifers. Royal Botanic Gardens, Kew, England.

FARRIS, J. S., M. KALLERSJO, A. G. KLUGE, and C. BULT. 1995. Testing significance of incongruence. Cladistics **10**:315–319.

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

FLORIN, R. 1963. The distribution of conifer and taxad genera in time and space. Acta Hort. Berg. **20**:121–312.

FRANKIS, M. P. 1988. Generic inter-relationships in Pinaceae. Notes RBG Edinb. **45**:527–548.

GAUSSEN, H. 1966. Les Gymnosperms actuelles et fossils. Trav. Lab. For. Toulouse Tome **2**:481–715.

GERNANDT, D. S., and A. LISTON. 1999. Internal transcribed spacer region evolution in *Larix* and *Pseudotsuga* (Pinaceae). Am. J. Bot. **86**:711–723.

GILLHAM, N. W. 1994. Organelle genes and genomes: transmission and compatibility of organelle genomes. Oxford University Press, New York.

GOLDMAN, N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. **36**:182–198.

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

HART, J. A. 1987. A cladistic analysis of conifers: preliminary results. J. Arn. Arb. **68**:269–307.

HIESEL, R., A. V. HAESELER, and A. BRENNICKE. 1994. Plant mitochondrial nucleic acid sequences as a tool for phylogenetic analysis. Proc. Natl. Acad. Sci. USA **91**:634–638.

HIPKINS, V. D., K. V. KRUTOVSKII, and S. H. STRAUSS. 1994. Organelle genomes in conifers: structure, evolution, and diversity. For. Genet. **1**:179–189.

JOHNSON, L. A., and D. E. SOLTIS. 1994. *matK* DNA sequences and phylogenetic reconstruction in Saxifragaceae sensu stricto. Syst. Bot. **19**:143–156.

———. 1995. Phylogenetic inference in Saxifragaceae sensu stricto and *Gilia* (Polemoniaceae) using *matK* sequences. Ann. Mo. Bot. Gard. **82**:149–175.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. N. MUNRO, ed. Mammalian protein metabolism. Academic Press, New York.

KIMURA, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

———. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. USA **78**:454–458.

KINLAW, C. S., and D. B. NEALE. 1997. Complex gene families in pine genomes. Trends Plant Sci. **2**:356–359.

LAROCHE, J., P. LI, L. MAGGIA, and J. BOUSQUET. 1997. Molecular evolution of angiosperm mitochondrial introns and exons. Proc. Natl. Acad. Sci. USA **94**:5722–5727.

LEE, D., M. ELLARD, L. A. WANNER, K. R. DAVIS, and C. J. DOUGLAS. 1995. The *Arabidopsis thaliana* 4-coumarate: CoA ligase (4CL) gene: stress and developmentally regulated expression and nucleotide sequence of its cDNA. Plant Mol. Biol. **28**:871–884.

LEPAGE, B. A., and J. F. BASINGER. 1995*a*. The evolutionary history of the genus *Larix* (Pinaceae). USDA For. Serv. Int. Res. Sta. GTR-INT **319**:19–29.

———. 1995*b*. Evolutionary history of the genus *Pseudolarix* Gordon (Pinaceae). Int. J. Plant Sci. **156**:910–950.

LIDHOLM, J., and P. GUSTAFSSON. 1991. A three-step model for the rearrangement of the chloroplast *trnK-pshA* region of the gymnosperm *Pinus contorta*. Nucleic Acids Res. **19**: 2881–2887.

LIN, J.-X., Y.-S. HU, and F. H. WANG. 1995. Wood and bark anatomy of *Nothotsuga* (Pinaceae). Ann. Mo. Bot. Gard. **82**:603–609.

LISTON, A., W. A. ROBINSON, J. M. OLIPHANT, and E. R. ALVAREZ-BUYLLA. 1996. Length variation in the nuclear ribosomal DNA internal transcribed spacer region of nonflowering seed plants. Syst. Bot. **21**:109–121.

MADDISON, W. P. 1997. Gene trees in species trees. Syst. Biol. **46**:523–536.

MARTIN, W., A. GIERL, and H. SAEDLER. 1989. Molecular evidence for pre-Cretaceous angiosperm origins. Nature **339**: 46–48.

MASON-GAMER, R. J., C. F. WEIL, and E. A. KELLOGG. 1998. Granule-bound starch synthase: structure, function, and phylogenetic utility. Mol. Biol. Evol. **15**:1658–1673.

MELCHIOR, H., and E. WERDERMANN. 1954. Engler, Syllabus der Pflanzenfamilien. 12th edition. Berlin.

MILLER, C. N. 1977. Mesozoic conifers. Bot. Rev. **43**:217–281.

———. 1988. The origin of modern conifer families. Pp.448–486 *in* C. B. BECK, ed. Origin and evolution of gymnosperms. Columbia University Press, New York.

MOGENSEN, H. L. 1996. The hows and ways of cytoplasmic inheritance in seed plants. Am. J. Bot. **83**:383–404.

MORTON, B. R., B. S. GAUT, and M. T. CLEGG. 1996. Evolution of alcohol dehydrogenase genes in the Palm and Grass families. Proc. Natl. Acad. Sci. USA **93**:11735–11739.

MURRAY, B. G. 1998. Nuclear DNA amounts in gymnosperms. Ann. Bot. **82**(Suppl. A):3–15.

MUSE, S. V., and B. S. WEIR. 1992. Testing for equality of evolutionary rates. Genetics **132**:269–276.

PAGE, C. N. 1988. New and maintained genera in the conifer families Podocarpaceae and Pinaceae. Notes RBG Edinb. **45**:377–395.

PALMER, J. D. 1992. Mitochondrial DNA in plant systematics: applications and limitations. Pp. 36–49 *in* P. S. SOLTIS, D. E. SOLTIS, and J. J. DOYLE, eds. Molecular systematics of plants. Chapman Hall, New York.

PERRY, D. J., and G. R. FURNIER. 1996. *Pinus banksiana* has at least seven expressed alcohol dehydrogenase genes in two linked groups. Proc. Natl. Acad. Sci. USA **93**:13020–13023.

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

PRICE, R. A., J. OLSEN-STOJKOVICH, and J. M. LOWENSTEIN. 1987. Relationships among the genera of Pinaceae: an immunological comparison. Syst. Bot. **12**:91–97.

QIU, Y.-L., and J. D. PALMER. 1999. Phylogeny of early land plants: insights from genes and genomes. Trends Plant Sci. **4**:26–30.

SANG, T., D. J. CRAWFORD, and T. F. STUESSY. 1997. Chloroplast phylogeny, reticulate evolution, and biogeography of *Paeonia* (Paeoniaceae). Am. J. Bot. **84**:1120–1136.

SANG, T., M. J. DONOGHUE, and D. ZHANG. 1997. Evolution of alcohol dehydrogenase genes in peonies (*Paeonia*): phylogenetic relationships of putative nonhybrid species. Mol. Biol. Evol. **14**:994–1007.

SAVARD, L., P. LI, S. H. STRAUSS, M. W. CHASE, M. MICHAUD, and J. BOUSQUET. 1994. Chloroplast and nuclear gene sequences indicated late Pennsylvanian time for the last common ancestor of extant seed plants. Proc. Natl. Acad. Sci. USA **91**:5163–5167.

SMALL, R. L., J. A. RYBURN, R. C. CRONN, T. SEELANAN, and J. F. WENDEL. 1998. The tortoise and the hare: choosing between noncoding plastome and nuclear *Adh* sequences for phylogeny reconstruction in a recently diverged plant group. Am. J. Bot. **85**:1301–1315.

SORHANNUS, U., and C. VAN BELL. 1999. Testing for equality of molecular evolutionary rates: a comparison between a relative-rate test and a likelihood ratio test. Mol. Biol. Evol. **16**:848–855.

STEELE, K. P., and R. VILGALYS. 1994. Phylogenetic analyses of Polemoniaceae using nucleotide sequences of the plastid gene *mat*K. Syst. Bot. **19**:126–142.

STEFANOVIC, S., M. JAGER, J. DEUTSCH, J. BROUTIN, and M. MASSELOT. 1998. Phylogenetic relationships of conifers inferred from partial 28S rRNA gene sequences. Am. J. Bot. **85**:688–697.

SWOFFORD, D. L. 1998. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer, Sunderland, Mass.

TEMPLETON, A. R. 1983. Phylogenetic inference from restriction endoclease cleavage site maps with particular reference to the evolution of humans and the apes. Evolution **37**:221–244.

THOMPSON, J. D., D. G. HIGGINS, and T. J. GIBSON. 1994. CLUSTAL W—improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

TSUDZUKI, J., K. NAKASHIMA, T. TSUDZUKI, J. HIRATSUKA, M. SHIBATA, T. WAKASUGI, and M. SUGIURA. 1992. Chloroplast DNA of black pine retains a residual inverted repeat lacking rRNA genes: nucleotide sequences of *trnQ*, *trnK*, *psbA*, *trnI* and *trnH* and the absence of *rps16*. Mol. Gen. Genet. **232**:206–214.

TSUMURA, Y., K. YOSHIMURA, N. TOMARU, and K. OHBA. 1995. Molecular phylogeny of conifers using RFLP analysis of PCR-amplified specific chloroplast genes. Theor. Appl. Genet. **91**:1222–1236.

VAN CAMPO-DUPLAN, M., and H. GAUSSEN. 1948. Sur quatre hybrides de genres chez les Abietinees. Trav. Lab. For. Toulouse Tome **24**:1–14.

VAN TIEGHEM, P. 1891. Structure et affinites des Abies et des genres les plus voisins. Bull. Soc. Bot. Fr. **38**:406–415.

WAGNER A., N. BLACKSTONE, P. CARTWRIGHT, M. DICK, B. MISOF, P. SNOW, G. P. WAGNER, J. BARTELS, M. MURTHA, and J. PENDLETON. 1994. Surveys of gene families using polymerase chain reaction: PCR selection and PCR drift. Syst. Biol. **43**:250–261.

WANG, X.-Q., Y. HAN, and D. Y. HONG. 1998*a*. A molecular systematic study of *Cathaya*, a relic genus of the Pinaceae in China. Plant Syst. Evol. **213**:165–172.

———. 1998*b*. PCR-RFLP analysis of the chloroplast gene *trnK* in the Pinaceae, with special reference to the systematic position of *Cathaya*. Isr. J. Plant Sci. **46**:265–271.

WENDEL, J. F., and J. J. DOYLE. 1998. Phylogenetic incongruence: window into genomes history and molecular evolution. Pp. 265–296 *in* D. E. SOLTIS, P. S. SOLTIS, and J. J. DOYLE, eds. Molecular systematics of plants, II: DNA sequencing. Kluwer, Boston.

WOLFE, K. H., M. GOUY, Y-W. YANG, P. M. SHARP, and W.-H. LI. 1989. Date of the monocot dicot divergence estimated from chloroplast DNA-sequence data. Proc. Natl. Acad. Sci. USA **86**:6201–6205.

ZHANG, X.-H., and V. L. CHIANG. 1997. Molecular cloning of 4-coumarate : coenzyme A ligase in loblolly pine and the roles of this enzyme in the biosynthesis of lignin in compression wood. Plant Physiol. **113**:65–74.

PAMELA SOLTIS, reviewing editor

Accepted January 25, 2000