Syst. Biol. 53(2):177–192, 2004

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150490423278

Recent Origin and Phylogenetic Utility of Divergent ITS Putative Pseudogenes: A Case Study from Naucleeae (Rubiaceae)

SYLVAIN G. RAZAFIMANDIMBISON,[1,2] ELIZABETH A. KELLOGG,[3] AND BIRGITTA BREMER[1,2]

[1]*Department of Systematic Botany, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18 D, SE-752 36 Uppsala, Sweden; E-mail: sylvain.razafimandimbison@ebc.uu.se (S.G.R.)*
[2]*The Bergius Foundation at the Royal Swedish Academy of Sciences, P.O. Box 50017, SE-104 05 Stockholm, Sweden*
[3]*University of Missouri–St. Louis, 8001 Natural Bridge Road, St. Louis, Missouri 63121, USA*

Abstract.—The internal transcribed spacer (ITS) of nuclear ribosomal DNA has been widely used by systematists for reconstructing phylogenies of closely related taxa. Although the occurrence of ITS putative pseudogenes is well documented for many groups of animals and plants, the potential utility of these pseudogenes in phylogenetic analyses has often been underestimated or even ignored in part because of deletions that make unambiguous alignment difficult. In addition, long branches often can lead to spurious relationships, particularly in parsimony analyses. We have discovered unusually high levels of ITS polymorphism (up to 30%, 40%, and 14%, respectively) in three tropical tree species of the coffee family (Rubiaceae), *Adinauclea fagifolia*, *Haldina cordifolia*, and *Mitragyna rubrostipulata*. Both secondary structure stability and patterns of nucleotide substitutions in a highly conserved region (5.8S gene) were used for distinguishing presumed functional sequences from putative pseudogenes. The combination of both criteria was the most powerful approach. The sequences from *A. fagifolia* appear to be a mix of functional genes and highly distinct putative pseudogenes, whereas those from *H. cordifolia* and *M. rubrostipulata* were identified as putative pseudogenes. We explored the potential utility of the identified putative pseudogenes in the phylogenetic analyses of Naucleeae sensu lato. Both Bayesian and parsimony trees identified the same monophyletic groups and indicated that the polymorphisms do not transcend species boundaries, implying that they do not predate the divergence of these three species. The resulting trees are similar to those produced by previous analyses of chloroplast genes. In contrast to results of previous studies therefore, divergent putative pseudogenes can be useful for phylogenetic analyses, especially when no sequences of their functional counterparts are available. Our studies clearly show that ITS polymorphism may not necessarily mislead phylogenetic inference. Despite using many different PCR conditions (different primers, higher denaturing temperatures, and absence or presence of DMSO and BSA-TMACl), we recovered only a few functional ITS copies from *A. fagifolia* and none from *H. cordifolia* and *M. rubrostipulata*, which suggests that PCR selection is occurring and/or the presumed functional alleles are located at minor loci (with few ribosomal DNA copies). [Concerted evolution; internal transcribed spacers; Naucleeae; nuclear ribosomal DNA; polymorphisms; pseudogenes.]

The internal transcribed spacer (ITS) of nuclear ribosomal DNA (nrDNA) is widely used in systematics, especially plant systematics, for reconstructing phylogenies of closely related taxa. For example, 2,900 ITS sequences (versus ca. 2,300 *rbc*L sequences) of angiosperm species had been submitted to GenBank as of 1998 (Hershkovitz et al., 1999:286). A survey by Álvarez and Wendel (2003) revealed that one third (34%) of the phylogenetic analyses published during the last 5 years were based exclusively on ITS sequences. The ITS region (ITS1, 5.8S gene, ITS2) is a component of a tandemly repeated multigene family that generally undergoes rapid concerted evolution (Zimmer et al., 1980) via unequal crossing over (Smith, 1976) and/or gene conversion (Arnheim, 1983). This region is phylogenetically useful partly because of sequence homogeneity among repeats within the same species (Hillis and Dixon, 1991; Hamby and Zimmer, 1992; Baldwin et al., 1995). Nevertheless, the degree of homogeneity in the tandemly repetitive sequence is the result of interplay between the rate of homogenization and the rate of new mutations (Ohta and Dover, 1983; Schlötterer and Tautz, 1994). Minimal polymorphism is expected within individuals or species as long as the rate of homogenization exceeds the rate of new mutations. In contrast, intraindividual or intraspecific variation is expected when the rate of homogenization is lower than the rate of new mutations (e.g., Schlötterer et al., 1994; Baldwin et al., 1995). Growing evidence from many angiosperm groups strongly suggests that the rate of homogenization and hence the level of polymorphism may differ among taxa, e.g., 0–7% in Winteraceae (Suh et al., 1993), 0–4.2% in the *Amelanchier* complex, Rosaceae (Campbell et al., 1997), 0–5.01% in *Aeschynanthus*, Gesneriaceae (Denduangboripant and Cronk, 2000), and 0.46–23.4% in *Aconitum*, Ranunculaceae (Kita and Ito, 2000). This paralogue diversity may include not only divergent functional alleles (e.g., this study) but also putative pseudogenes or nonfunctional alleles (e.g., Buckler and Holtsford, 1996a, 1996b; Muir et al., 2001; this study) and recombinants (e.g., Buckler et al., 1997). Although complete homogenization of repeats is advantageous for phylogeneticists, intermediate levels of concerted evolution make reliable phylogenetic reconstruction difficult (e.g., Sanderson and Doyle, 1992). Putative pseudogenes are sometimes excluded a priori from phylogenetic analyses (e.g., Yang et al., 1999) because they could experience long-branch attraction (Felsenstein, 1978) as a result of parallel substitutions.

Despite increasing evidence of its occurrence in both animals and plants, ITS polymorphism has not been evaluated intentionally (Hershkovitz et al., 1999). It has been detected and subsequently investigated only when amplification of genomic DNA produced multiple bands or when direct sequencing of PCR products of apparently uniform bands consistently yielded multiple sequence signals. The present study was motivated by the discovery of high levels of ITS polymorphism in three tree species, *Adinauclea fagifolia*, *Haldina cordifolia*, and

Mitragyna rubrostipulata, of the tribe Naucleeae sensu lato (Razafimandimbison and Bremer, 2001, 2002). Both intraindividual and intraspecific ITS polymorphisms were associated with the persistence of highly divergent putative pseudogenes. However, no such polymorphism was found in the remaining Naucleeae taxa investigated.

Naucleeae sensu lato belongs to the subfamily Cinchonoideae (of the coffee family, Rubiaceae) and is a well-defined monophyletic group whose members can be easily diagnosed by their numerous flowers arranged in globose inflorescences and epigynous floral nectaries deeply embedded in hypanthia. The tribe comprises ca. 179 species and 26 genera of trees, shrubs, and woody climbers and has its center of diversity in tropical Asia (128 species), mostly Southeast Asia, followed by Madagascar (24 species), the African mainland (22 species), and Central, North, and South America together (5 species) (Razafimandimbison and Bremer, 2001, 2002). Both *Adinauclea* and *Haldina* are monotypic and exclusively Asian. *Adinauclea* is restricted to the lowland rainforests of the island of Sulawesi (Indonesia) and is only known from a few old herbarium specimens and is therefore considered rare. The material used for this study was collected from one individual growing at the Bogor Botanical Garden (Indonesia). *Haldina* is restricted to dry habitats of India through Vietnam and Sri Lanka. *Mitragyna rubrostipulata*, an obligate rheophyte of East Africa, is one of the four African species of the Afro-Asian genus *Mitragyna* sensu lato (Ridsdale, 1978; Razafimandimbison and Bremer, 2002). *Mitragyna* belongs to the monotypic subtribe Mitragyninae sensu Ridsdale (1978), whereas both *Adinauclea* and *Haldina* are members of the Asian subtribe Adininae sensu Razafimandimbison and Bremer (2002). Many members of Naucleeae sensu lato are economically important. *Haldina cordifolia*, *Mitragyna ciliata*, *M. stipulosa*, and *Nauclea diderichii* (Bilinga) all produce high-quality woods. *Pausinystalia johimbe* (yohimbe) is a well-known aphrodisiac legally sold in drugstores. The North American *Cephalanthus occidentalis* (buttonbush) and *Haldina cordifolia* are commonly cultivated in many botanical gardens throughout the world.

We characterized all identified paralogues of *A. fagifolia*, *H. cordifolia*, and *M. rubrostipulata* by examining their DNA secondary structure stabilities and patterns of nucleotide substitutions in a highly conserved region (5.8S gene). These data allowed us to hypothesize that some of the sequences are putative pseudogenes. We then explored the potential phylogenetic utility of putative pseudogenes using Naucleeae sensu lato as an example. Unlike in previous such studies (e.g., Buckler and Holtsford, 1996b; Kita and Ito, 2000; Mayol and Rossello, 2001; Muir et al., 2001), we found that the putative pseudogenes are closely related to the functional alleles in the same species. This finding suggests that these pseudogenes are of recent origin, i.e., they are not evolutionary relicts, and therefore they do not interfere with our ability to infer phylogenetic relationships. We also investigated the effects of PCR conditions on products recovered by PCR assays.

Materials and Methods

Taxon Sampling

We investigated 23 of the 179 described species of Naucleeae sensu lato (Razafimandimbison and Bremer, 2002) and two outgroup taxa, *Cinchona pubescens* and *Exostema lineatum* (Table 1), which also belong to Rubiaceae. The choice of outgroup taxa was based on findings from our earlier studies (Razafimandimbison and Bremer, 2001, 2002). One individual, with the exception of *Breonia decaryana* and *Haldina cordifolia* from which two individuals were investigated, represented each taxon.

Experiment 1: 95°C, ITS Primers

DNA extraction and amplification.—Total DNA, extracted from leaves dried in silica gel (Chase and Hills, 1991) and/or herbarium material, was isolated following the miniprep procedure of Saghai-Maroof et al. (1984) as modified by Doyle and Doyle (1987). DNA templates were amplified using the primer pair ITS-I/ITS4 (Table 2). Two PCR additives were used to determine whether they affected the products recovered by PCR. The first PCR profile included 1 μl of 10% dimethylsulfoxide (DMSO), whereas the second one included 0.5 μl of 1% of bovine serum albumin (BSA) and 5 μl of 0.1 M tetramethylammonium chloride (TMACl) together in all 50-μl PCRs. PCR amplifications, performed in a MJ Research machine (Peltier Thermal Cycler) and/or an Eppendorf Mastercycle gradient (Eppendorf), began with an initial melting phase of 2 min at 95°C, followed by 35 cycles of 60 sec at 95°C, 90 sec at 55°C, and 90 sec at 72°C, and a final extension phase of 7 min at 72°C. Hereinafter, we refer to this experiment as the 95°/ITS experiment. Amplification of *M. rubrostipulata* always resulted in two distinct bands: one longer band, verified by sequencing, of 607–612 base pairs (bp) and the other shorter band of ca. 516 bp. We then loaded the entire reaction volume onto a 1% agarose gel, and both longer and shorter bands were excised and purified separately with glassmilk (BIO 101) according to the manufacturer's instructions before reamplification. Amplification of the remaining taxa produced single bands, which were similar in length to the longer band of *M. rubrostipulata*. In all PCRs, one reaction was run with water instead of DNA as a negative control to check for contamination.

Cloning and sequencing reactions.—PCR products were cloned using either the Promega cloning kit (Promega Corp., A13809) or the TOPO TA cloning kit (Invitrogen). The Promega cloning kit was used for the purified PCR products of *A. fagifolia* and *H. cordifolia* amplified with DMSO, which were ligated into the pGEM-T easy vector systems. The transformation mixture was incubated in SOC medium at 37°C with agitation and plated on LB agar with ampicillin (50 μg/ml), IPTG, and X-Gal (25 μg). Four white colonies from the cloning reaction of *A. fagifolia* and individual A of *H. cordifolia* were selected for growth. Plasmid DNA was isolated according to an alkaline lysis miniprep protocol (Sambrook et al., 1989).

TABLE 1. List of taxa used in this study, paralogues identified, and accession numbers.

Taxa[a]	No. clones	Paralogues identified[b]	Accession nos.
Polymorphic taxa[c]			
Experiment 1: 95°/ITS primers			
Adinauclea fagifolia	22		
Amplified with BSA-TMACl	18	Ad1–14-BT	AJ492632–45
Amplified with DMSO	4	Ad12-D	AJ492643
		Ad15–17-D	AJ492646–8
Haldina cordifolia	13		
Individual A amplified with DMSO	4	HAL1–3-D	AJ492622–4
Individual B amplified with BSA-TMACl	9	HAL4–10-BT	AJ492625–31
Mitragyna rubrostipulata	15		
Amplified with BSA-TMACl	15	MIT1–14-BT	AJ492608–21
Experiment 2: 97°/rRNA primers			
Adinauclea fagifolia	9		
Amplified with DMSO	9	Ad18–26-D	AJ605955–63
Haldina cordifolia	17		
Individual B amplified with BSA-TMACl	5	HAL11–15-BT	AJ605964–68
Individual B amplified with DMSO	12	HAL16–27-D	AJ605969–80
Mitragyna rubrostipulata	14		
Amplified with BSA-TMACl	4	MIT15–18-BT	AJ605981–84
Amplified with DMSO	10	MIT19–28-D	AJ605985–93
Remaining taxa			
Breonadia salicina			AJ346857
Breonia decaryana			AJ346859
Burttdavya nyassica			AJ346863
Cephalanthus occidentalis			AJ346883
Cephalanthus salicifolius			AJ346886
Gyrostipula foveolata			AJ346867
Janotia macrostipula			AJ346869
Metadina trichotoma			AJ346871
Mitragyna diversifolia			AJ346872
Mitragyna inermis			AJ346873
Mitragyna stipulosa			AJ346868
Myrmeconauclea strigosa			AJ346875
Nauclea diderrichii			AJ346855
Neolamarckia cadamba			AJ346878
Neonauclea forsteri			AJ346880
Pausinystalia macroceras			AJ346890
Pseudocinchona mayumbensis			AJ346921
Sarcocephalus latifolius			AJ346899
Uncaria africana			AJ414545
Uncaria guianensis			AJ414546
Cinchona pubescens			AJ224838
Exostema lineatum			AJ346902

[a]Voucher references: *Cinchona pubescens*, Andreasen et al. (1999); *H. Cordifolia* individual A, Lorence 750166.001, PTBG, Hawaii; all others, Razafimandimbison and Bremer (2002).

[b]BT = BSA-TMACl; D = DMSO. All identified paralogues of *A. fagifolia*, *H. cordifolia*, and *M. rubrostipulata*, with the exception of five presumed functional sequences (Ad3, Ad9, Ad11, Ad14, and Ad26), are putative pseudogenes (Ψ).

[c]We recovered Ad3, Ad12, Ad19, HAL3, HAL8, HAL10, and MIT13 twice and Ad11 three times.

The ITS region was digested from the plasmid with *Eco*RI according to PEG preparation protocol (Perkin-Elmer, Bulletin no. 18) prior to sequencing using the plasmid primers T7 and sp6 and the internal primers ITS2 and ITS3B (Table 2).

The TOPO TA cloning kit was used for *A. fagifolia*, individual B of *H. cordifolia*, and *M. rubrostipulata* amplified with BSA-TMACl. This kit used unpurified PCR-amplified DNA, the TOPO vector, and a vial of One Shot chemically competent *Escherichia coli* according to the manufacturer's instructions. Eighteen, 9, and 15 white colonies from the cloning reactions of *A. fagifolia*, individual B of *H. cordifolia*, and *M. rubrostipulata*, respectively, were screened and amplified with the T7 and M13R universal primers (Table 2). Sequencing reactions were prepared using four primers: ITS-I, ITS2, ITS3B, and ITS4 (Table 2).

Experiment 2: 97°C, Ribosomal Primers

Based on the results with the 95°/ITS experiment, we performed additional PCR amplifications of *A. fagifolia*, *H. cordifolia*, and *M. rubrostipulata* using the same PCR conditions but at a denaturing temperature of 97°C and with a different pair of external primers (P17/26S-82R; Table 2) priming just outside of the 5′ end of ITS1 and the 3′ end of ITS2, respectively. Hereinafter, based on the PCR conditions we refer to this experiment as the 97°/rRNA experiment. We were unable to investigate individual A of *H. cordifolia* because of lack of material. All three

TABLE 2. Primers used in this study.

Primers	Sequence, 5′ 3′	Reference
Forward		
ITS-I	GTC CAC TGA ACC TTA TCA TTT AG	Urbatsch et al. (2000)
P17	CTA CCG ATT GAA TGG TCC GGT GAA	Pop and Oxelman (2001)
ITS2	CGT AAC AAG GTT TCC GTA GG	Baum et al. (1994)
T7[a]	AAT ACG CTC ACT ATA G	
Reverse		
ITS3B	GCA TCG ATG AAG AAC GTA GC	Baum et al. (1994)
ITS4	TCC TC GCT TAT TGA TAT GC	White et al. (1990)
26S-82R	TCC CGG TTC GCT CGC CGT TAC TA	Pop and Oxelman (2001)
SP6[a]	GTA TTA GGT GAC ACT ATA G	
M13R[a]	CAG GAA ACA GCT ATG AC	

[a]Primer included in cloning kits and used for the present study.

taxa were amplified with DMSO, and *H. cordifolia* and *M. rubrostipulata* were also amplified with BSA-TMACl. For *M. rubrostipulata*, we directly cloned its PCR products without excising the two distinct bands to determine whether sequences with intermediate lengths (between the two distinct bands) were present but not visible on our gels. Using the TOPO TA cloning kit, we screened and cloned 9, 12, and 10 colonies of *A. fagifolia*, *H. cordifolia*, and *M. rubrostipulata*, respectively, from the unpurified PCR products amplified with DMSO. Similarly, five and four colonies of *H. cordifolia* and *M. rubrostipulata*, respectively, from the unpurified PCR products amplified with BSA-TMACl were screened and cloned (see Table 1, experiment 2).

Data Analysis

We performed BLAST searches using the ITS clonal sequences of *A. fagifolia*, *H. cordifolia*, and *M. rubrostipulata* to determine whether the sequences obtained were from fungal endophytes or other contaminants. Sequence data were aligned using CLUSTAL W (Thompson et al., 1994) to produce an initial alignment. This process was followed by manual alignment using Se-Al (Rambaut, 1996).

Both secondary structure stability and patterns of nucleotide substitutions in an otherwise highly conserved region (5.8S gene) were used as criteria for distinguishing presumed functional sequences from putative pseudogenes. Minimum-energy secondary structures of both ITS1 and ITS2 (5.8S excluded) of each divergent sequence of *A. fagifolia*, *H. cordifolia*, and *M. rubrostipulata* were estimated with MFold, a web-based program of the Macfarlane Burnet Centre (http://mfold.burnet.edu.au), using the default temperature of 37°C. MFold was developed by Zuker (1989) to predict optimal and suboptimal secondary structures for RNA or DNA molecules. The HYPERMUT Program Package (Rose and Korber, 2000), a web-based interface (http://www.hiv.lanl.gov/content/hiv-db/HYPERMUT/hypermut.html), was used to document the patterns of nucleotide substitutions in all identified paralogues of each polymorphic species relative to the reference sequence, *Cephalanthus salicifolius*. In our previous studies (Razafimandimbison and Bremer, 2001, 2002), *Cephalanthus* was consistently placed as sister to the remaining members of Naucleeae. This program was originally designed to study the sequence evolution of HIV, with a very high level of G → A mutations. We assumed that all differences arose from a single substitution, neglecting the possibility of multiple substitutions. The program summarizes all substitutions (A ↔ C, A ↔ G, A ↔ T, C ↔ G, C ↔ T, and G ↔ T) observed in each sequence being compared with the reference sequence and highlights their physical locations along the sequences.

We used PAUP* 4.0b8b (Swofford, 2000) to compute both distances between all pairs of divergent sequences (excluding all ambiguously aligned sites) and the frequency of G and C bases of all identified sequences of *A. fagifolia*, *H. cordifolia*, and *M. rubrostipulata*. We adopted the HKY85 model of substitution (Hasegawa et al., 1985) drawn from a gamma distribution (HKY85+Γ) to estimate sequence pairwise distances. The Bayesian estimate of α for the entire ITS data (1.543) was used for computing all HKY85+Γ-corrected distances. This model was selected as the best model from a comparison of 56 models using the Akaike information criterion (Akaike, 1974) as implemented in Modeltest 3.06 (Posada and Crandall, 1998). It allows both unequal base frequencies and transitions and transversions to occur at different rates.

To investigate the phylogenetic utility of putative pseudogenes, we performed both Bayesian and parsimony phylogenetic analyses including all divergent paralogous sequences generated from the 95°/ITS and the 97°/rRNA experiments: 26 sequences of *A. fagifolia*, 27 sequences of *H. cordifolia*, 28 sequences of *M. rubrostipulata*, 20 sequences of the remaining investigated Naucleeae taxa, and two outgroup sequences, for a total of 103 sequences. Identical sequences were represented by single sequences in all phylogenetic analyses, and gaps were treated as missing data.

Bayesian phylogenetic analyses were conducted with the computer program MrBayes 3.0b4 (Huelsenbeck and Ronquist, 2001) using the HKY85+Γ substitution model and 2 million Markov chain Monte Carlo (MCMC) generations. The temperature of the chains and other parameters were left at default values. To evaluate whether the chain ran for enough generations, we performed up to four independent runs, each with a different random starting tree and sampling every 1,000 trees. After the end of the four runs, we determined and compared the burn-ins based on graphs with generation on the *x*-axis versus log probability of observing the data on the *y*-axis. We also verified that means and variances of

the model likelihood after burn-ins were similar. The remaining trees saved (after excluding the burn-ins) were subsequently used to construct Bayesian majority rule consensus trees with the help of PAUP*. We additionally compared the majority-rule consensus trees from the independent runs to verify that their topologies and clade credibility values were similar (Huelsenbeck et al., 2002). Checks of these additional parameters all confirmed convergence after the initially determined burn-in phase. All saved trees from the independent runs (burn-ins excluded) were pooled for a consensus tree. To produce a quantitative estimate of among-site rate heterogeneity, we conducted additional Bayesian analyses to estimate the shape parameter (α) of the gamma distribution for all ITS sequences of each polymorphic species separately. The smaller the value of α, the more extreme the gamma distribution and hence rate heterogeneity. A series of Bayesian analyses also were carried out to address additional questions. We determined whether using paralogous sequences recovered by different PCR conditions in analyses affected phylogenetic results. Two Bayesian analyses were performed, one including only sequences of *A. fagifolia*, *H. cordifolia*, and *M. rubrostipulata* amplified with DMSO plus the 20 sequences from the remaining Naucleeae taxa, and the other including all sequences of the polymorphic species amplified with BSA-TMACl plus the 20 other Naucleeae sequences. We also assessed whether incomplete concerted evolution had an effect on phylogenetic accuracy by comparing the trees from Bayesian analyses, in which each polymorphic species was represented by one of its identified paralogue sequences, with the tree from the Bayesian analysis with all identified paralogues included. We also examined whether the inclusion of a large number of putative pseudogenes relative to the number of functional sequences in analyses partly contributed to the increase level of homoplasy. We kept the total number of analyzed sequences constant but changed the proportion of putative pseudogene sequences to functional sequences or vice versa.

Because the current implementation of the parsimony method does not adequately deal with among-site rate variation (Yang, 1996), we intentionally performed an equally weighted parsimony analysis to test whether this violation led to incorrect phylogenies. The following settings were used: heuristic searches with uninformative and ambiguously aligned sites all excluded, the MULPARS option on, tree bisection–reconnection branch swapping, 5,000 random sequence additions, and unordered and equally weighted characters. We calculated the consistency index (CI; Kluge and Farris, 1969) with uninformative characters excluded and the retention index (RI; Farris, 1989). Ten thousand jackknife (JK) replicates using a deletion frequency of 37%, emulate "jac" (Farris et al., 1996) resampling option on, MULPARS option off, nearest-neighbor interchanges (NNI) branch swapping, and five random sequence additions were performed to assess relative support for retained clades.

RESULTS

Amplification and Cloning

Direct sequencing of the purified PCR products of *A. fagifolia*, individuals A and B of *H. cordifolia*, and *M. rubrostipulata* consistently produced multiple sequence signals, indicating the presence of intraindividual and intraspecific polymorphisms. Cloning of the PCR products of these three taxa revealed highly divergent paralogues. Direct sequencing of purified products from all other taxa yielded unambiguous sequences. To verify that these PCR products contained no cryptic variation, we cloned two or three colonies from each and found no intraindividual or intraspecific polymorphisms. Despite deletions in 26 putative pseudogenes, the sequences were easy to align with the published ITS Naucleeae sequences included in the present study (see Table 1). BLAST searches using the cloned ITS sequences of *A. fagifolia*, *H. cordifolia*, and *M. rubrostipulata* revealed that none of these paralogue sequences were identical to the published ITS Naucleeae sequences, but the latter sequences were among the top BLAST hits. This finding was corroborated by the results of our phylogenetic analyses presented here. Therefore, the possibility that these sequences are contaminants can be dismissed. The number of sequences sampled from each polymorphic species is summarized in Table 1. A total of 50 and 40 ITS clonal sequences of the polymorphic species were generated from the 95°/ITS and the 97°/rRNA experiments, respectively.

Sequence Length and Divergence of Sequence Pairs

The lengths of the nonaligned ITS-region sequences ranged from 569 to 613 bp, 523 to 610 bp, and 516 to 612 bp in *A. fagifolia*, *H. cordifolia*, and *M. rubrostipulata*, respectively. From the 97°/rRNA experiment, we identified five sequences of *M. rubrostipulata* (MIT16–17-BT, MIT21-D, MIT23-D, and MIT27-D) with lengths (584–599 bp) falling between the shorter (516 bp) and the longer (607–612 bp) distinct bands. Lengths of the ITS region varied from 604 to 610 bp for the remaining Naucleeae and from 593 to 595 bp for the two outgroup taxa. In *A. fagifolia*, *H. cordifolia*, and *M. rubrostipulata*, the HKY85+Γ-corrected distances for the entire ITS region within each taxon ranged from 0.35% to 29.66%, 0.17% to 40.38%, and 6.50% to 14.16%, respectively. The distance for *A. fagifolia* and *H. cordifolia* were higher than those for any interspecific comparison among the remaining Naucleeae (1.62–12.56%) and comparable to those between ingroup and outgroup taxa (17.83–42.10%).

Sequence Stability, GC Content, and Substitution Patterns

For each sequence, the MFold program computed up to 16 sequence foldings; the estimates with the lowest energy folds were the most energetically favored or optimal structures, which we report here. For all identified sequences recovered from the three polymorphic taxa, we plotted the estimated values of the free energies

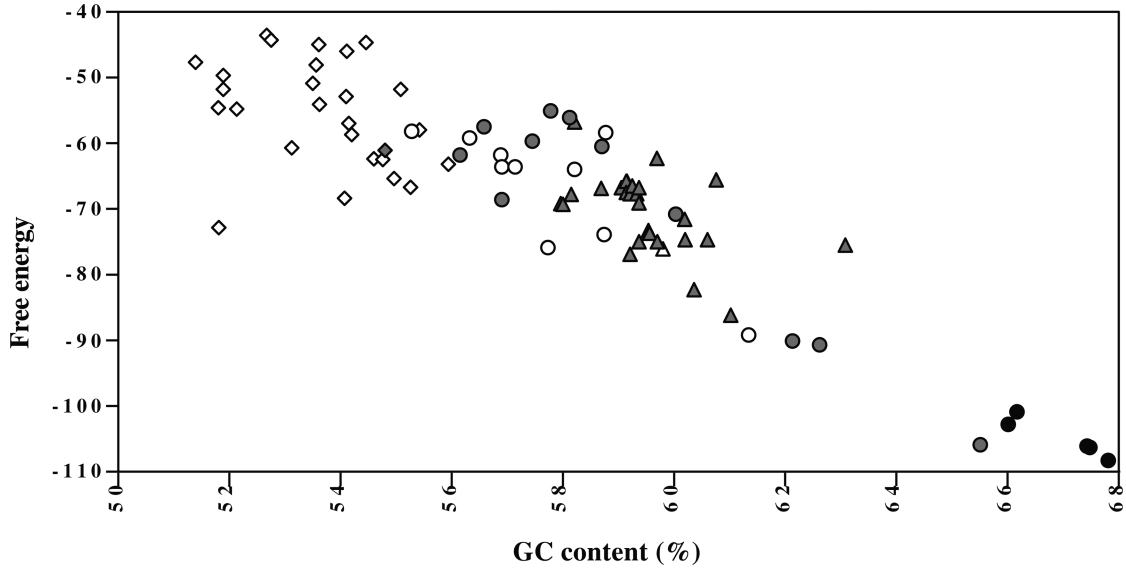

FIGURE 1. Correlation between calculated free energies and GC contents in the divergent sequences of *A. fagifolia* (○), *H. cordifolia* (●), and *M. rubrostipulata* (●). Shaded: putative pseudogenes with 2–9 mutations in the 5.8S region; Open: putative pseudogenes with >10 mutations in the 5.8S region; Solid: presumed functional sequences with single mutation in the 5.8S region.

(kcal/mol; representing the secondary structure stabilities) of the investigated sequences against those of GC contents (%) (Fig. 1). In general, the sequences with less stable secondary structures (high free energies) and those with low GC content clustered together. The estimated free energies of the putative pseudogenes of *A. fagifolia*, *H. cordifolia*, and *M. rubrostipulata* ranged from −55.1 to −90.7 (−105.9), −43.6 to −72.6, and −56.8 to −86.2 kcal/mol, respectively. Their GC contents ranged from 55.25% to 62.62% (65.51%), 51.39% to 55.94%, and 57.96% to 61.02% (63.08%), respectively. The range of the estimated free energies and GC content varied from −100.9 to −108.3 kcal/mol and 66.1% to 67.48%, respectively, for the presumed functional sequences of *A. fagifolia*. Among the low-stability sequences, those from *H. cordifolia* were the least stable (Fig. 1). All sequences with low-stability secondary structures and at least two mutations within the 5.8S region were considered putative pseudogenes (ψ), and those sequences of *A. fagifolia* with high-stability secondary structures and at most one mutation within the 5.8S region were considered functional sequences.

The physical locations of all mutations in each sequence as related to the reference sequence are shown in Figures 2–4. Within the 5.8S gene, five sequences of *A. fagifolia* (Ad3-BT, Ad9-BT, Ad11-BT, Ad14-BT, and Ad26-D) had only a single mutation (Fig. 2). All of these sequences except Ad26-D were produced from the 95°/ITS experiment. For all putative pseudogenes of *A. fagifolia*, *H. cordifolia*, and *M. rubrostipulata* generated from both the 95°/ITS and the 97°/rRNA experiments, the number of mutations within the 5.8S region ranged from 3 to 20, 12 to 27, and 2 to 14, respectively. Only one putative pseudogene, MIT21ψ, has two mutations within the 5.8S gene.

Phylogenetic Analysis

The sequence alignment required a number of indels ranging from 2 to 95 bp; these resulted in ca. 3% of the cells in the matrix being occupied by gaps. The aligned matrix was 635 bp ALIGN_000653 and contained 478 (ca. 77%) parsimony-informative characters. The alignment was ambiguous for 44 sites (7% of the total), which were excluded from all analyses. Twenty-eight (63.63%) of these 44 sites were parsimony informative. Of the 591 remaining aligned sites, 450 (76.14%) were parsimony informative. The Bayesian majority rule consensus tree pooled from the Bayesian trees (burn-ins excluded) from the four independent runs and the parsimony strict consensus tree generated from 22,323 most-parsimonious trees (each with length = 3,085, CI = 0.264, and RI = 0.532) are shown in Figures 5 and 6, respectively (S1041). With *Cinchona pubescens* and/or *Exostema lineatum* as outgroup taxa, both the Bayesian and the parsimony trees identified the same large monophyletic groups. Both trees were consistent with (1) Cephalanthinae and a strongly supported clade with *Neonauclea forsteri* and *Myrmeconauclea strigosa* as sister taxa; (2) *Metadina trichotoma* unresolved; (3) the monophyly of the subtribes Cephalanthinae, Mitragyninae, Uncarinae, Naucleinae, Corynantheinae, Breoniinae, and Mitragyninae (all sensu Razafimandimbison and Bremer, 2001, 2002); (4) all identified clonal sequences from *A. fagifolia*, *H. cordifolia*, and *M. rubrostipulata* forming separate monophyletic groups; (5) Naucleinae, Corynantheinae, and Breoniinae forming a poorly supported monophyletic group; (6) all presumed functional sequences of *A. fagifolia* together with one putative pseudogene (Ad21Ψ-D) forming a strongly supported clade; and (7) the phylogenetic relationships among the subtribes unresolved.

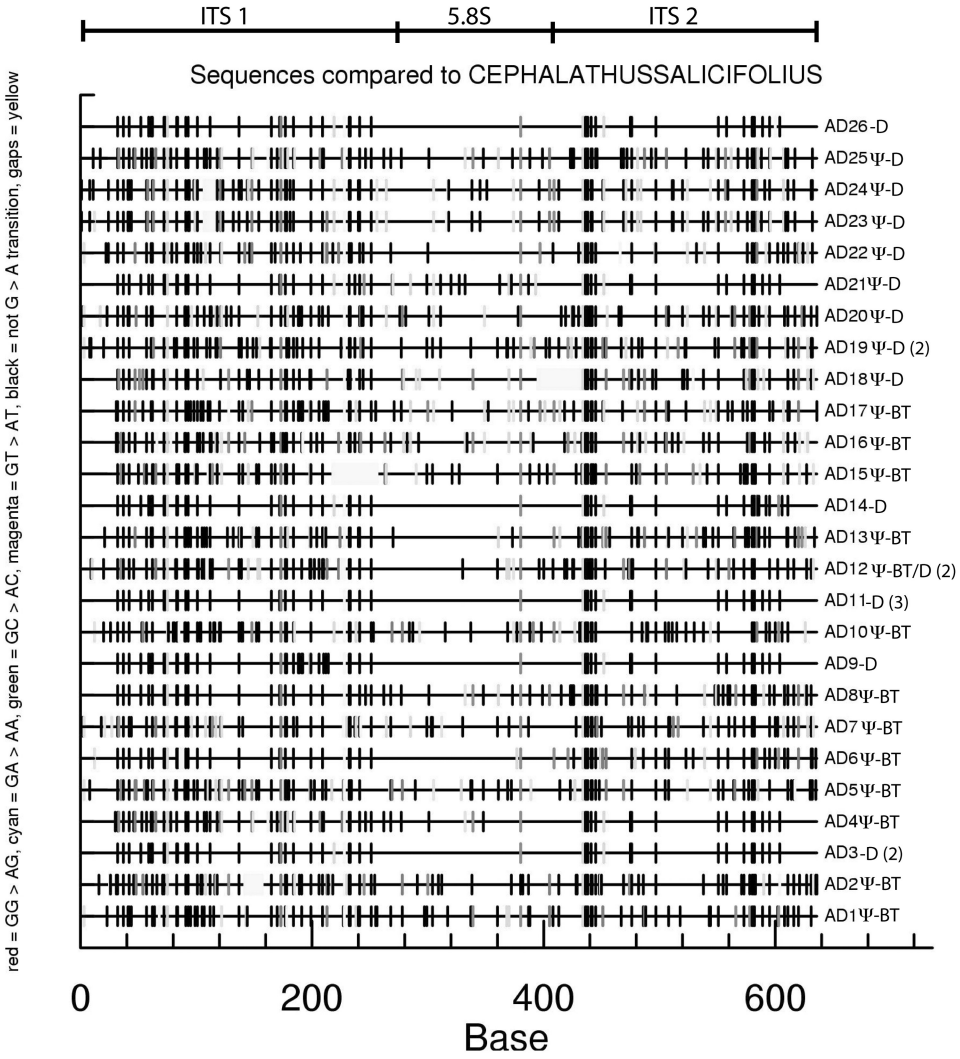

FIGURE 2. Distribution of substitutions across the ITS region of all the identified ITS paralogues of *A. fagifolia* from the 95°/ITS and 97°/rRNA experiments. BT and D stand for amplifications with BSA-TMACl and DMSO, respectively. ψ = putative pseudogene. Numbers in parentheses correspond to the numbers of identical sequences.

The Bayesian and parsimony trees (Figs. 5, 6), however, presented some poorly supported differences. The clade containing Cephalanthinae and a subclade of *Neonauclea forsteri* and *Myrmeconauclea strigosa* was resolved as basal in the Bayesian tree (Fig. 5). However, in the parsimony tree (Fig. 6) it was placed and left unresolved in a very large polytomy containing all the remaining investigated Naucleeae taxa. Although Uncarinae was placed at the base in the parsimony tree, it was resolved with poor support (posterior probability [PP] = 50) as sister to a clade of Naucleinae, Corynantheinae, and Breoniinae.

The Bayesian trees (results not presented) generated from separate analyses including all paralogue sequences amplified with DMSO and BSA-TMACl, respectively, were similar to the trees shown in Figures 5 and 6, with only slight differences in support. When the three polymorphic species were represented by only one of their paralogue sequences, the overall tree topologies and support (not presented) were also similar to those shown in Figures 5 and 6. When the total number of analyzed sequences was kept constant but the proportion of putative pseudogene sequences relative to the functional sequences was increased, the level of homoplasy gradually increased. Conversely, the level of homoplasy gradually diminished with the gradual increase of the number of presumed functional sequences.

Bayesian Estimates of Rate Variation Across Sites

The shape parameter of the gamma distribution, a measure of among-site rate variation, was 1.543 for the entire matrix. When all identified paralogues of *A. fagifolia*, *H. cordifolia*, and *M. rubrostipulata* were analyzed separately, the estimated values of α (burn-ins excluded) were 1.14, 1.058, and 0.87, respectively,

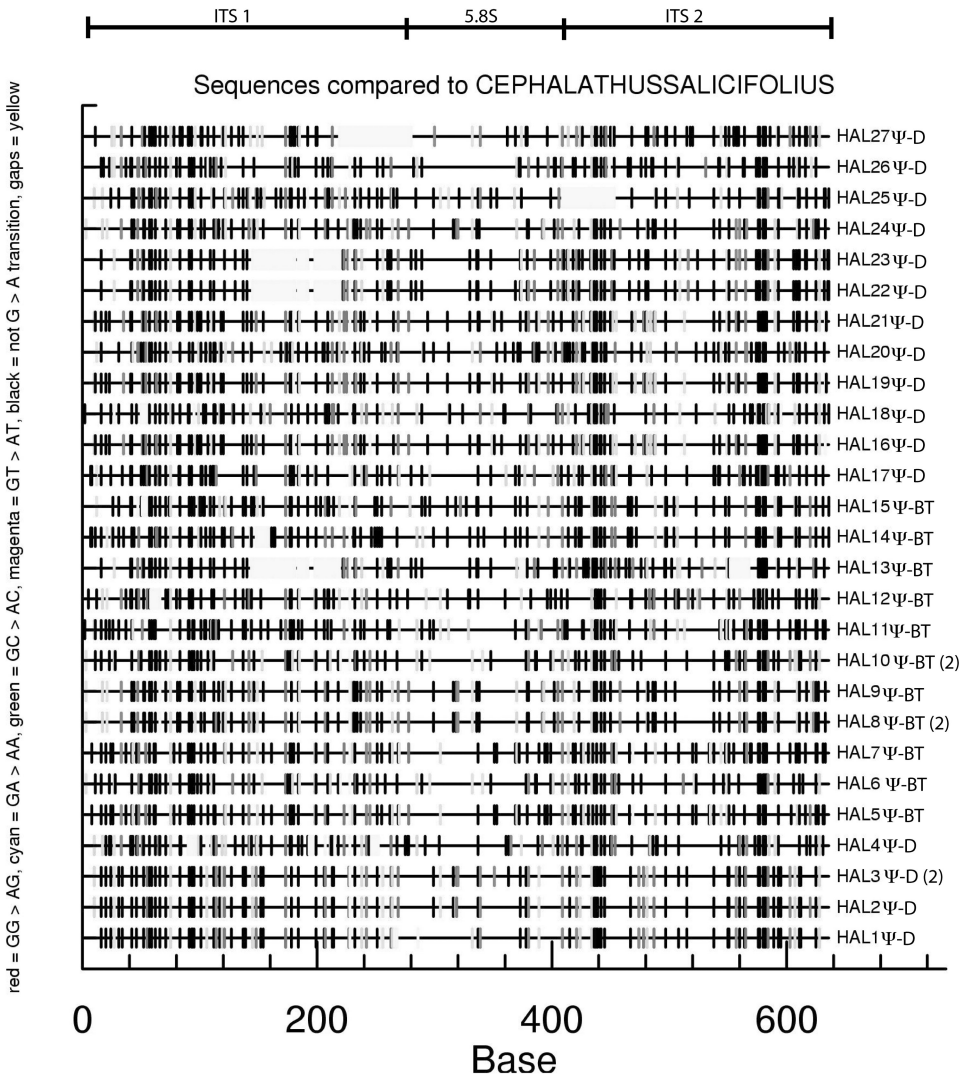

FIGURE 3. Distribution of substitutions across the ITS region of all the identified ITS paralogues of *H. cordifolia* from the 95°/ITS and 97°/rRNA experiments. BT and D stand for amplifications with BSA-TMACl and DMSO, respectively. ψ = putative pseudogene. Numbers in parentheses correspond to the number of identical sequences.

suggesting moderate among-site rate variation. When analyzed separately, the estimated value of α was 1.449 for all the putative pseudogenes of *A. fagifolia*, indicating low rate variation, and 24.79 for all functional sequences of *A. fagifolia*, indicating unusually low among-site rate variation.

DISCUSSION

Paralogue Identification

In general, secondary structure stability and patterns of nucleotide substitutions in the 5.8S gene suggest identical sets of sequences as putative pseudogenes and presumed functional sequences. The combination of both criteria appears to be the most powerful approach for distinguishing putative pseudogenes from presumed functional sequences. Most of the sequences of *A. fagifolia* (23 of 31; Figs. 1, 2) and all the sequences of *H. cordifolia* ($n = 30$, Figs. 1, 3), and *M. rubrostipulata* ($n = 29$; Figs. 1, 4) are putative pseudogenes. These sequences all have lower GC content, corresponding to the relatively high frequency of substitutions to A and T, and higher free energies, i.e., less stable secondary structures (Fig. 1) than the functional sequences identified from *A. fagifolia*. They have at least two mutations in the 5.8S region (Figs. 1–4). We identified five sequences of *A. fagifolia* (Ad3-BT, Ad9-BT, Ad11-BT, Ad14-BT, and Ad26-D) as functional alleles because they have higher GC contents (66.1–67.48%), lower free energies, i.e., more stable secondary structures, and only a single mutation in the 5.8S region (Figs. 1, 2). Their pairwise divergences range from 0.35% to 4.69%.

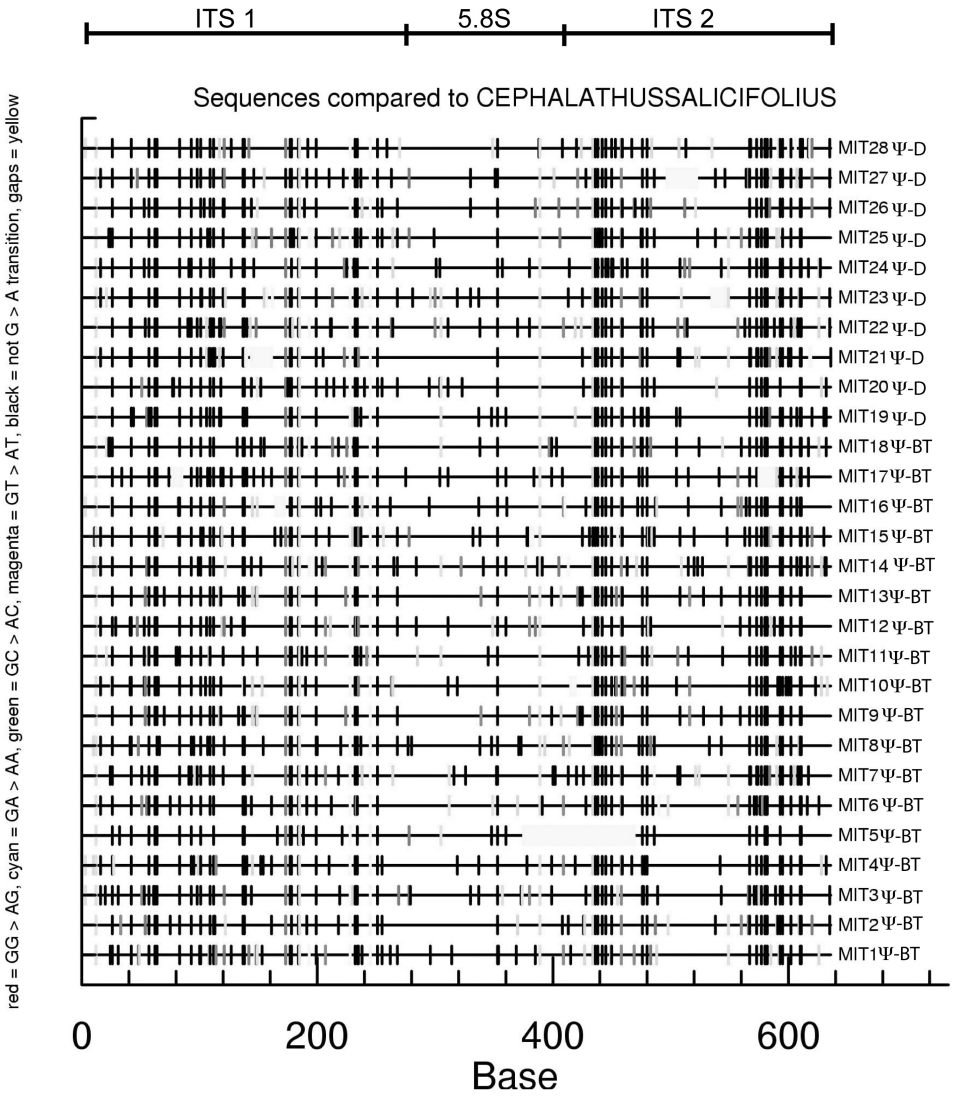

FIGURE 4. Distribution of substitutions across the ITS region of all the identified ITS paralogues of *M. rubrostipulata* from the 95°/ITS and 97°/rRNA experiments. BT and D stand for amplifications with BSA-TMACl and DMSO, respectively. ψ = putative pseudogene. Numbers in parentheses correspond to the number of identical sequences.

Hershkovitz et al. (1999) argued that nucleotide substitutions within the 5.8S region appear to be more reliable indicators for distinguishing functional sequences from putative pseudogenes because nrDNA function might be disabled by even one mutation without substantially affecting secondary structure. Based on secondary structure stability alone, one could argue that Ad4, Ad6, and Ad8 represent functional sequences because they still have fairly high GC content (61.34–62.62%) and stable secondary structures (−90.7 to −89.2 kcal/mol) compared with all other putative pseudogenes with lower GC content (51.39–61.02%) and less stable secondary structure (−43.6 to −86.2 kcal/mol). However, we consider these three sequences to be putative pseudogenes because their 5.8S regions have several mutations (4–17; Fig. 2). Despite its high GC content (63.08%; Fig. 1), we consider MIT5Ψ-D a pseudogene because it has a deletion of 95 bp (Fig. 4), an unstable secondary structure (−75.5 kcal/mol; Fig. 1), and five mutations in the 5.8 region (Fig. 4). All other truncated alleles (Figs. 2–4) are considered pseudogenes.

When estimating minimum free energy secondary structures of ITS sequences, some workers (e.g., Buckler et al., 1997) included only ITS1 and ITS2 and excluded the 5.8S gene; others (e.g., Hartmann et al., 2001), however, considered only the 5.8S gene and excluded ITS1 and ITS2. For our study, we followed the approach of Buckler et al. (1997), which produced conflicting results for Ad21. The free energies of Ad21 without and with the 5.8S gene are −105.9 and −122.4 kcal/mol, respectively. The free energy of Ad21 without the 5.8S gene (−105.9) is much lower (more stable) than the estimated free energy

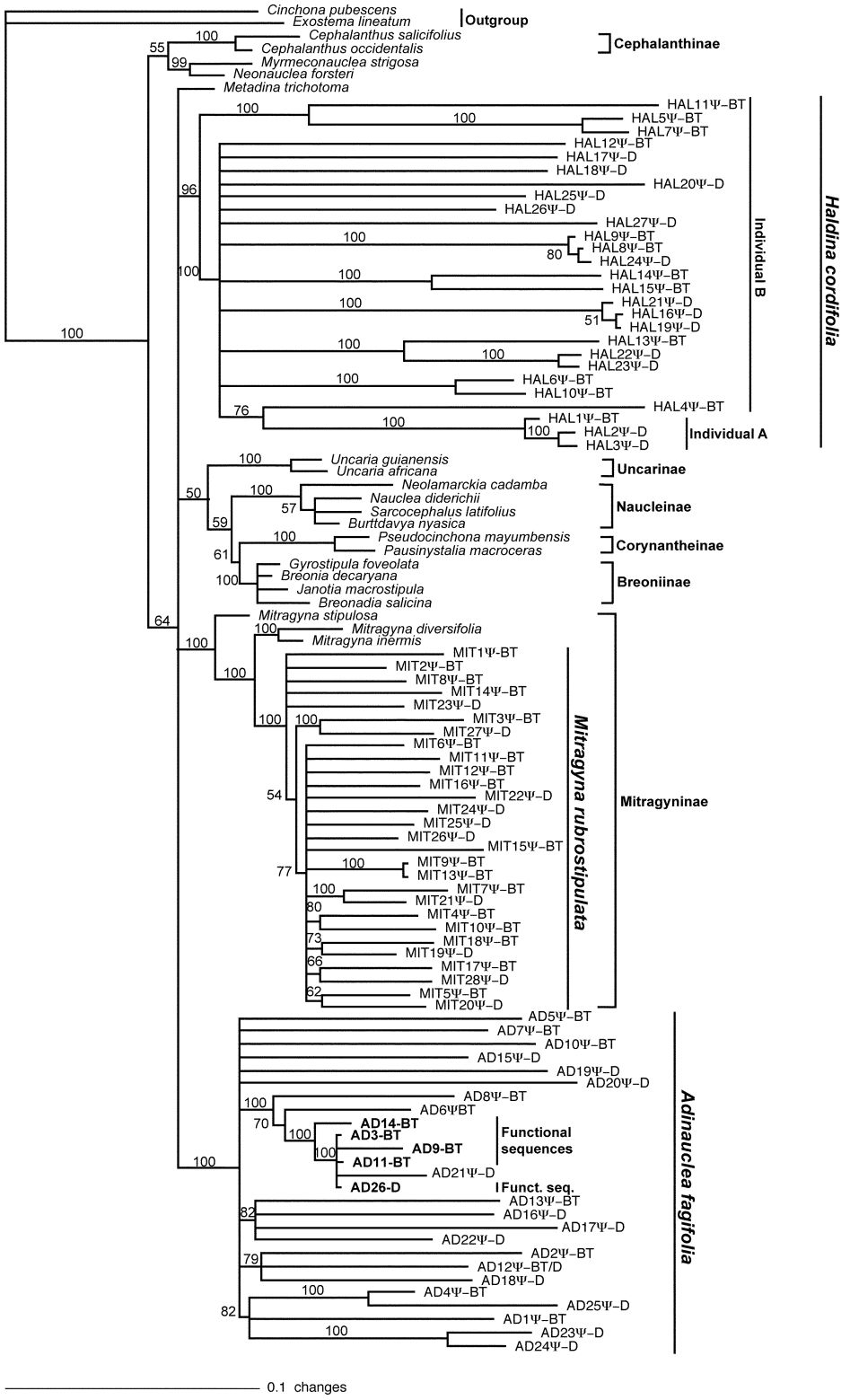

FIGURE 5. Fifty percent Bayesian majority rule consensus tree under the HKY85+Γ model of substitution from a 2-million MCMC generation analysis, showing mean branch lengths. Numbers on internodes indicate posterior probabilities. Functional sequences of *A. fagifolia* are bold. ψ = putative pseudogene. Vertical bars delimit outgroup taxa, functional sequences of *A. fagifolia*, and the three polymorphic species. Brackets correspond to the subtribes.

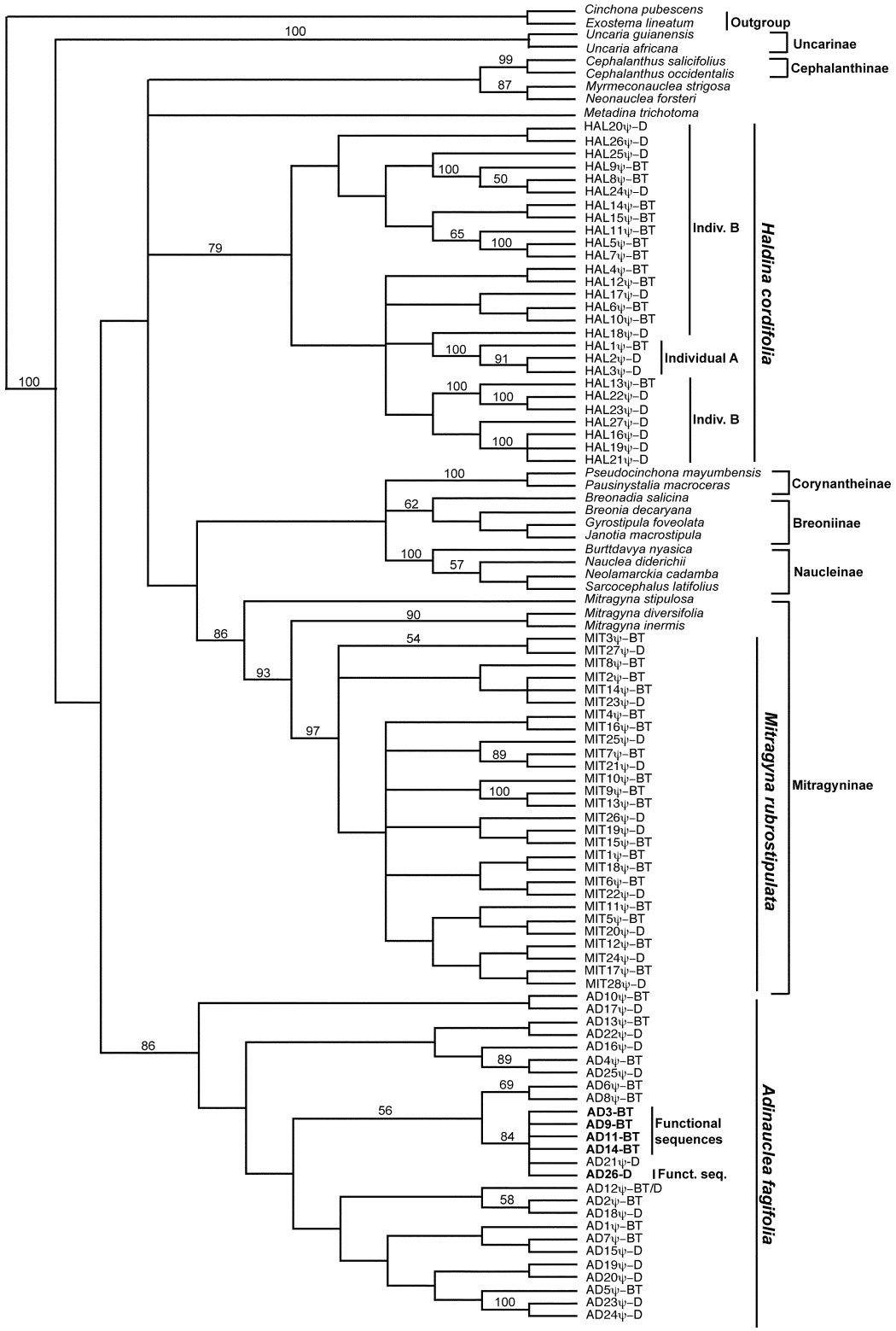

FIGURE 6. Parsimony strict consensus tree of 22,323 most-parsimonious trees of ITS data (length = 3,085; CI = 0.264; RI = 0.532). Numbers on internodes are JK support (>50%) for that internode. ψ = putative pseudogene. Vertical bars delimit the outgroup taxa and the three polymorphic taxa. Brackets correspond to the subtribes.

(without 5.8S region) of the presumed functional sequence Ad14 (−100.9 kcal/mol), indicating that Ad21 should also be considered a functional sequence. However, the value with the 5.8S region included (−122.4) is much higher (less stable) than all estimated free energies of the identified functional sequences of *Adinauclea* (5.8S included), which range from −125.8 to −133.5 kcal/mol (more stable), indicating that Ad21 represents a putative pseudogene. We conclude that Ad21 is indeed a putative pseudogene and its 5.8S gene actually contains 12 mutations (Figs. 1, 2).

Phylogenetic Utility of ITS Putative Pseudogenes

In this study, we established the prevalence of highly divergent pseudogenes in *A. fagifolia*, *H. cordifolia*, and *M. rubrostipulata* of Naucleeae sensu lato. The range of pairwise divergence values between sequences of the remaining investigated taxa (1.62–12.56%) is comparable to that found within *M. rubrostipulata* (up to 14%) but is much lower than that found within *A. fagifolia* (up to 30%) and *H. cordifolia* (up to 40%). Such intraindividual and intraspecific polymorphisms together with the high homoplasy level (CI = 0.264) could potentially reduce the phylogenetic utility of the ITS region. In both Bayesian and parsimony analyses, however, all investigated paralogue sequences of the three polymorphic species form separate monophyletic groups. Also, they diagnose the same monophyletic groups (Breoniinae, Cephalanthinae, Corynantheinae, Naucleinae, Mitragyninae, and Uncarinae), which also have been identified in studies of other genes (Razafimandimbison and Bremer, 2001, 2002). Thus, the analyses presented (including 76 highly divergent ITS pseudogenes of the polymorphic species) verify the reliability of the phylogenetic relationships and the stability of the intratribal classifications of Naucleeae addressed in our earlier study. The analyses also show that A. *fagifolia*, *H. cordifolia*, and *M. rubrostipulata* are all distinct lineages. This conclusion is consistent with morphological data (Ridsdale, 1978; Razafimandimbison and Bremer, 2002), and their current taxonomic status should be maintained.

Because divergent paralogous sequences of the same individuals or species are always together in the trees (Figs. 5, 6), there is no indication in our data that the observed polymorphisms predate the divergence of these taxa. Therefore, despite considerable intraindividual and intraspecific variation in these three polymorphic species it is still possible to use these ITS sequence data to reconstruct phylogenies of Naucleeae.

Buckler and Holtsford (1996a, 1996b) and Buckler et al. (1997) reported the occurrence of a few putative pseudogenes that had escaped homogenization since before the divergence of *Zea* (Poaceae), and interpreted these pseudogenes as evolutionary relicts. They suggested that these types of pseudogenes could be used as better outgroups than sister species and could even be more useful for groups without closely related extant taxa. This study clearly shows that putative pseudogenes (e.g., those from *H. cordifolia* and *M. rubrostipulata*) can be useful for phylogenetic analyses when no sequences of their functional counterparts are available.

The present study includes 19 of the 44 and 22 of the 52 ITS sequences used by Razafimandimbison and Bremer (2001 and 2002, respectively) in their studies of Naucleeae sensu lato. Yet the number of parsimony-informative characters from the ITS data used in this study (450) is almost 2.5 times higher than that found by Razafimandimbison and Bremer (2001). This number (450) is also much higher than that from both the combined molecular (ITS-*rbc*L-*trn*T-F) data (381) and the combined molecular and morphological data (429), respectively, used by Razafimandimbison and Bremer (2002). However, only ca. 31% of all possible internodes are resolved (with JK > 50%) when all ITS paralogue sequences are included in the parsimony analysis (Fig. 6). Most of the internodes within *Adinauclea*, *Haldina*, and *M. rubrostipulata* clades are collapsed or have JK support of <50%. In contrast, 65.90% and 69.81% of all potential internodes are resolved in the parsimony analyses of Razafimandimbison and Bremer (2001 and 2002, respectively). Thus, some of the many variable and parsimony-informative sites (450 of the 591 aligned sites) in our data set represent homoplasy rather than useful variation. Our analyses also show that inclusion of many divergent putative pseudogenes in phylogenetic analyses has partly contributed to the increased homoplasy (CI = 0.264 vs. 0.482 and 0.466 for Razafimandimbison and Bremer, 2001, 2002, respectively) in our ITS data and has tremendously increased computational time. Such highly divergent sequences could also experience long-branch attraction (Felsenstein, 1978) due to parallel substitutions in two or more taxa erroneously identified as synapomorphies. At least in this study, this is not the case; both our Bayesian and parsimony analyses identified the same monophyletic groups. Therefore, exclusion of putative pseudogenes a priori from phylogenetic analyses, as done by some workers (e.g., Yang et al., 1999), may not be necessary.

Sanderson and Doyle (1992), in their simulation study, demonstrated that intermediate levels of concerted evolution can blur the distinction between paralogous and orthologous copies without nullifying it; they argued that this blurring makes reliable phylogenetic reconstruction difficult. Our results, however, indicate that the situation might not be as bleak as they suggested. The overall topologies and the level of resolution of the trees (with all paralogue sequences included) shown in Figures 5 and 6 are similar to the trees (not shown) generated from the same data set but with the three polymorphic species represented by only one of their identified paralogues. These results were also corroborated by Razafimandimbison and Bremer (2001, 2002); both *H. cordifolia* and *M. rubrostipulata* were similarly represented by one of their paralogue sequences. Thus, inclusion of the highly divergent paralogues in analyses does not interfere with our ability to infer phylogenetic relationships. In this case, incomplete concerted evolution may not necessarily affect phylogenetic accuracy.

Estimation of the Alpha Parameter of the Gamma Distribution

In general, rate variation among sites in putative pseudogenes is expected to be low because functional constraints are equally relaxed over all sites. In contrast, among-site rate variation in functional sequences is supposed to be higher than that of putative pseudogenes because all sites are under functional constraints (Yang, 1996). Therefore, the very low estimate of rate of variation among sites ($\alpha = 24.79$) from the five functional sequences of *A. fagifolia* (Ad3-BT, Ad9-BT, Ad11-BT, Ad14-BT, and Ad26-D) is surprising. Three of these five sequences, Ad3-BT, Ad11-BT, and Ad26-D, are almost identical, differing by only three changes. Only 26 (all within ITS1 and ITS2) of 591 aligned sites (ambiguously aligned sites excluded) are variable among the five sequences. Thus, the overall rate of variation is low enough that variation among sites might not be detected.

The estimates of branch lengths and the shape parameter of the gamma distribution (α) have been shown to be affected by the tree construction methods and models (Yang, 1996). For example, estimates of α using parsimony tree topology are all much higher (data not presented) than the Bayesian estimates for the entire matrix and all paralogues of *A. fagifolia*, *H. cordifolia*, and *M. rubrostipulata* separately.

Effects of Among-Site Rate Variation and High Levels of Homoplasy on Parsimony-Based Phylogenetic Inference

Simulation studies (e.g., Kuhner and Felsenstein, 1994) have shown that the performance of the parsimony method in recovering the "correct" phylogeny deteriorates significantly when among-site rate variation exists, whereas both maximum likelihood and Bayesian methods (model-based methods) can deal with unequal rates of substitutions among sites (Huelsenbeck, 1995). In the present study, among-site rate variation ($\alpha = 1.543$) is unusually and unexpectedly low, which may explain why the parsimony tree (Fig. 6) identifies the same large monophyletic groups as the Bayesian tree (Fig. 5).

Maximum parsimony can give misleading results when homoplasy is common or concentrated in a particular part of the tree (e.g., Huelsenbeck, 1995). Álvarez and Wendel (2003) argued that misleading results may be obtained when homoplastic characters are distributed such that they resolve as synapomorphies. Our studies, however, show that this is not necessarily the case; the tree topologies estimated from maximum parsimony (Fig. 6) appear not to be misled by the high level of homoplasy (CI = 0.264) in our data set because the putative pseudogenes still have enough synapomorphies to hold them together.

Correlates of High Levels of Polymorphism

Levels of intraindividual and intraspecific ITS polymorphims reported here are among the highest known from angiosperms. The high within-individual polymorphism found in *M. rubrostipulata* was not observed in the other three investigated *Mitragyna* species (*M. diversifolia*, *M. inermis*, and *M. stipulosa*). We were unable to investigate the possible occurrence of such polymorphism in other individuals of *A. fagifolia* and *M. rubrostipulata* because of lack of material.

Several biological phenomena such as high ploidy level (Suh et al., 1993), allopolyploidy (e.g., Karvonen and Savolainen, 1993), long generation time (Sang et al., 1995), agamospermy or asexual seed production (e.g., Campbell et al., 1997), presence of pseudogenes (e.g., Buckler and Holtsford, 1996a, 1996b; Kita and Ito, 2000; Muir et al., 2001; Álvarez and Wendel, 2003), and a large number of nucleolar organizer regions (NORs) (Bobola et al., 1992; Karvonen and Savolainen, 1993) all have been suggested to be responsible for high levels of ITS polymorphisms. Based on secondary structure stability and patterns of nucleotide substitutions in the 5.8S region, we concluded that the high levels of intraindividual and intraspecific polymorphisms observed in *A. fagifolia*, *H. cordifolia*, and *M. rubrostipulata* are associated with the persistence of highly divergent putative pseudogenes.

Polymorphisms in *A. fagifolia*, *H. cordifolia*, and *M. rubrostipulata* do not appear to be the result of polyploidy per se. Except for the octoploid *Nauclea orientalis* (Kiehn and Lorence, 1996), all taxa investigated so far in Naucleeae (e.g., *Mitragyna parviflora* [Mehra and Gill, 1968], *M. rubrostipulata* [Kiehn, 1985], *Uncaria elliptica* [Kiehn, 1986], *Adina pilulifera*, *A. rubella*, *Breonia chinensis*, *Cephalanthus glabratus*, *C. occidentalis*, *Haldina cordifolia*, *Nauclea diderrichii*, *N. xanthoxylon*, *Neolamarckia cadamba*, *Sarcocephalus latifolius*, and *S. pobeguinii* [Kiehn, 1995]) are tetraploid. If polymorphism were a simple result of polyploidy, we would find it in all species of Naucleeae. However, we detected no intraindividual polymorphism in the two individuals of *Nauclea orientalis*, even though it is octoploid, nor did we find any in the remaining taxa. This result is consistent with small amounts of DNA for the haploid chromosome sets in *N. orientalis*, an indication of an already homogenized, old polyploid (Kiehn, 1986, 1995). If polymorphism were the result of hybridization and/or allopolyploidy, there would be some evidence of this from comparison with chloroplast gene trees. However, phylogenetic analyses of Naucleeae based on the uniparentally inherited *rbc*L (Razafimandimbison and Bremer, 2001) and *trn*T-F data sets (Razafimandimbison and Bremer, 2002) suggest the same placement for *M. rubrostipulata* as analyses based on ITS sequences. The *rbc*L and *trn*T-F sequence data (Razafimandimbison and Bremer, 2002) were inconclusive for assessing the exact position of *A. fagifolia* and *H. cordifolia* in Naucleeae because of lack of resolution. Concerted evolution can homogenize ITS in the direction of either parent (e.g., Wendel et al., 1995), so we cannot rule out allopolyploidy, but there is no evidence for it.

Long generation time has also been suggested as a mechanism that might retard concerted evolution (e.g., Sang et al., 1995). All members of Naucleeae except the woody climbers *Uncaria* are shrubs or trees, and these

have relatively long generation times. Although we do not have any detailed information on the generation times of *A. fagifolia*, *H. cordifolia*, and *M. rubrostipulata*, we found no evidence that their life histories are different from those of the rest of Naucleeae.

Campbell et al. (1997) suggested that agamospermy could be responsible for the high level of intraindividual ITS polymorphism found in the *Amelanchier* agamic complex (Rosaceae). Meiosis is absent in agamospermous plants; wherein megasporocytes degenerate and nearby somatic cells develop mitotically into chromosomally unreduced megagametophytes (Campbell and Wright, 1996). As a result, the heteroduplex molecules required for efficient molecular turnover and concerted evolution are missing. However, no agamospermous taxa in Naucleeae have been reported. All investigated members of Naucleeae (e.g., *Cephalanthus*, *Haldina*, *Neonauclea*, *Sarcocephalus*, and *Uncaria*) are protandrous, present their own pollen secondarily on the immature stigmas (Puff et al., 1996), and are self-incompatible (e.g., Imbert and Richards, 1993). Thus, the high level of within-individual polymorphism observed in *A. fagifolia*, *H. cordifolia*, and *M. rubrostipulata* may not be associated with agamospermy. However, facultative autonomous apomictic seed formation has recently been identified in the genus *Coprosma* of the subfamily Rubioideae of the family Rubiaceae (Heenan et al., 2003).

Karvonen and Savolainen (1993) suggested that within-individual polymorphism observed in Scots pine is associated with the presence of a large number of NORs; they found a total of eight NORs per haploid genome. Concerted evolution occurs more quickly within than between loci (e.g., Ohta and Dover, 1983). Ribosomal DNA (rDNA) gene organization, exact number of NOR loci, and rDNA repeat numbers in *A. fagifolia*, *H. cordifolia*, and *M. rubrostipulata* and the remaining Naucleeae genera are currently unknown. The high levels of polymorphisms found in these three taxa could indicate a large number of NORs. Fluorescent in situ hybridization (Maluszynska and Heslop-Harrison, 1991) could be used to elucidate the organization of rDNA within the genome of these taxa. Restriction mapping (e.g., Bobola et al., 1992) and Southern blotting (e.g., Copenhaver and Pikaard, 1996) could be utilized for quantifying rDNA tandem repeats. Unfortunately, we were unable to try any of these three techniques because of lack of appropriate plant materials. For the same reason, we were unable to check for transcription of the putative pseudogenes.

We do not have data on the flanking 18S or 26S genes for any of our ITS sequences. It would be interesting (although time consuming) to determine whether the putatively nonfunctional ITS sequences are flanked by equally nonfunctional ribosomal genes.

Effects of PCR Conditions on PCR Products

Evidence from several studies (e.g., Ritland et al., 1993; Wagner et al., 1994; Buckler et al., 1997) has indicated that PCR conditions can strongly favor particular gene copies such as pseudogenes or organismal contaminants, so that functional alleles are not obtained. Adding up to 10% DMSO to PCRs improves amplification of DNA with complex secondary structure (Shen and Hohn, 1992) and increases the specificity of primer matching (Filichkin and Gelvin, 1992). Buckler et al. (1997) showed that only alleles with high stability were amplified with DMSO, whereas both alleles with highly stable and less stable conformations were amplified without DMSO. BSA stabilizes *Taq* polymerases and also is useful when attempting to amplify ancient DNA, or templates that contain PCR inhibitors such as melanin (Kreader, 1996). Low concentrations of TMACl increase the yield and specificity of PCR (Chevet et al., 1995). Despite using many different PCR conditions (different primers, higher denaturing temperatures, and absence or presence of DMSO and BSA-TMACl) in our experiments, we recovered only a few presumed functional ITS copies from *A. fagifolia* and none from *H. cordifolia* and *M. rubrostipulata*. Thus, our findings from the 95°/ITS and the 97°/rRNA experiments are consistent, but these findings are inconsistent with the conclusions of Buckler et al. (1997). PCR selection (Wagner et al., 1994) could explain why more low-GC pseudogenes (23) than high-GC presumed functional copies (5) were amplified in *A. fagifolia*. Alternatively, the functional ITS copies could reside at a minor functional locus (with few rDNA copies) and the putative pseudogenes could be located at one or more major inactive loci. In that case, it would make sense that the number of functional copies remains lower than that of the putative pseudogenes after amplification. The same arguments could be used to explain the patterns observed in *H. cordifolia* and *M. rubrostipulata*. Furthermore, all five identified functional sequences (Ad3-BT, Ad9-BT, Ad11-BT, Ad14-BT, and Ad26-D) of *A. fagifolia* are placed in a strongly supported (PP = 100, JK = 84) monophyletic group with one putative pseudogene, AD21Ψ-D. This result would be expected if these presumed functional sequences represented a minor functional locus.

Concerted evolution is predicted to occur more quickly within than between rDNA loci (Ohta and Dover, 1983). Because highly divergent paralogues appear to be common in *A. fagifolia*, *H. cordifolia*, and *M. rubrostipulata*, the homogenization process may well have a limited genomic scope and rate both within and between rDNA loci. The divergence between the identified functional sequences of *A. fagifolia* ranges from 0.35% to 4.69%, consistent with this hypothesis. This range is comparable to the within-individual polymorphisms reported from the *Amelanchier* complex, Rosaceae (0–4.2%; Campbell et al., 1997), *Aeschynanthus*, Gesneriaceae (0–5.01%; Denduangboripant and Cronk, 2000), and Winteraceae (0–7%; Suh et al., 1993). We recovered Ad11-BT three times and Ad3-BT twice (see Table 1); the pairwise divergence between these two sequences is only 0.35%, suggesting that concerted evolution, although not homogenizing all sites, is operating.

PCR Artifacts

Some of the substitutions obtained from PCRs and cloning may be the result of *Taq* polymerase errors, estimated to range between 0.06 and 0.0006 bases per 100 bp of ITS clone (Abramson, 1995). This error level is far lower than the observed level of polymorphisms in *A. fagifolia* (up to 30%), *H. cordifolia* (up to 40%), and *M. rubrostipulata* (up to 14%). The same PCR profiles and cloning techniques were used for the remaining taxa in this study, but no such polymorphisms were detected in the other species. We are not aware of any study that addresses *Taq* fidelity in the presence of additives such as DMSO, BSA, and TMACl, but we found no indication in our results that these additives elevate errors.

In this study, we established the prevalence of putative pseudogenes in plants of the coffee family based on estimates of secondary structure stability, percentage GC content, and pattern of nucleotide substitutions. Unlike in previous studies, however, this high level of polymorphism did not compromise our ability to reconstruct phylogenies. The observed polymorphisms do not transcend species boundaries, indicating that they do not predate the divergence of *A. fagifolia*, *H. cordifolia*, and *M. rubrostipulata*. The ITS data sets used in the present study are appropriate for assessing phylogenetic relationships within Naucleeae despite the long branches of the putative pseudogenes and the high level of homoplasy. It would be interesting to determine whether the same levels of ITS polymorphism occur in different individuals of *A. fagifolia* and *M. rubrostipulata*.

ACKNOWLEDGMENTS

We thank Colin Ridsdale (National Herbarium of Leiden, The Netherlands) and David Lorence (National Tropical Botanical Garden, Hawaii) for providing material for *A. fagifolia* and *H. cordifolia*, respectively; Nahid Heidari and Edith Barkhordarian for help with sequencing; and Bengt Oxelman, Katarina Andreasen, Magnus Pop, and Henrik Lantz for their comments on an earlier version of the manuscript. We particularly thank Chris Simon, François Lutzoni, Bruce Baldwin, Kerry O'Donnell, and David Posada for providing constructive comments that greatly improved the manuscript. S.G.R. thanks Michael Möller and Ed Buckler for sharing their knowledge on ITS polymorphism, and Johan Nylander for helping with MrBayes. Financial support to B.B. for a postdoctoral position for S.G.R. was provided by the Swedish Research Council. Parts of this research were conducted at the E. Desmond Lee Molecular Systematics Laboratory of the University of Missouri–St. Louis as part of the Ph.D. program of S.G.R., who was also supported by the Andrew Mellon Foundation and Missouri Botanical Garden.

Note: A recent paper (Bailey et al., 2003. Characterization of angiosperm nrDNA polymorphism, paralogy, and pseudogenes. Mol. Phylogenet. Evol. 29:435–455) has reached conclusion similar to ours; in particular they show that ITS pseudogenes should not be excluded a priori from phylogenetic analyses.

REFERENCES

Abramson, R. D. 1995. Thermostable DNA polymerases. Pages 121–129 *in* PCR strategies (M. A. Innis, D. H. Gelfand, and J. J. Sninsky, eds.). Academic Press, San Diego.

Akaike, H. 1974. A new look at the statistical model identification. IEEE Trans. Auto. Cont. AC-19:716–723.

Álvarez, I., and J. F. Wendel. 2003. Ribosomal ITS sequences and plant phylogenetic inference. Mol. Phylogenet. Evol. 29:417–434.

Andreasen, K., B. Baldwin, and B. Bremer. 1999. Phylogenetic utility of the nuclear rDNA ITS region in subfamily Ixoroideae (Rubiaceae): Comparisons with cpDNA *rbc*L sequence data. Plant Syst. Evol. 217:119–135.

Arnheim, N. 1983. Concerted evolution of multiple gene families. Pages 38–61 *in* Evolution of genes and proteins (M. Nei and R. K. Koehn, eds.). Sinauer, Sunderland, Massachusetts.

Baldwin, B. G., M. J. Sanderson, M. J. Porter, M. F. Wojciechowski, C. S. Campbell, and M. J. Donoghue. 1995. The ITS region of nuclear ribosomal DNA: A valuable source of evidence on angiosperm phylogeny. Ann. Mo. Bot. Gard. 82:247–277.

Baum, D. A., K. J. Sytsma, and P. C. Hoch. 1994. A phylogenetic analysis of *Epilobium* (Onagraceae) based on nuclear ribosomal DNA sequences. Syst. Bot. 19:363–388.

Bobola, M. S., D. E. Smith, and A. S. Klein. 1992. Five major nuclear ribosomal DNA repeats represent a large and variable fraction of the genomic DNA of *Picea rubens* and *P. mariana*. Mol. Biol. Evol. 13:612–622.

Buckler, E. S., A. Ippolito, and T. P. Holtsford. 1997. The evolution of ribosomal DNA: Divergent paralogues and phylogenetic implications. Genetics 145:821–832.

Buckler, E. S., IV, and T. P. Holtsford. 1996a. *Zea* systematics: Ribosomal ITS evidence. Mol. Biol. Evol. 13:612–622.

Buckler, E. S., IV, and T. P. Holtsford. 1996b. *Zea* ribosomal repeat evolution and substitution patterns. Mol. Biol. Evol. 13:623–632.

Campbell, C. S., M. F. Wojciechowski, B. G. Baldwin, L. A. Alice, and M. Donoghue. 1997. Persistent nuclear ribosomal DNA sequence polymorphism in the *Amelanchier* agamic complex (Rosaceae). Mol. Biol. Evol. 14:81–90.

Campbell, C. S., and W. A. Wright. 1996. Apomixis, hybridization, and taxonomic complexity in eastern North American *Amelanchier* (Rosaceae). Folia Geobot. Phytotaxon. 55:345–354.

Chase, M. W., and H. H. Hills. 1991. Silica gel: An ideal material for preservation of leaf samples for DNA studies. Taxon 40:215–220.

Chevet, E., G. Lemaître, and M. D. Katinka. 1995. Low concentrations of tetramethylammonium chloride increase yield and specificity of PCR. Nucleic Acids Res. 23:3343–3344.

Copenhaver, G. P., and C. S. Pikaard. 1996. RFLP and physical mapping with a rDNA-specific endonuclease reveals that nucleolus organizer regions of *Arabidopsis thaliana* adjoin the telomeres on chromosomes 2 and 4. Plant J. 9:259–272.

Denduangboripant, J., and Q. C. B. Cronk. 2000. High intraindividual variation in internal transcribed spacer sequences in *Aeschynanthus* (Gesneriaceae): Implications for phylogenetics. Proc. R. Soc. Lond. B 267:1407–1415.

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.

Farris, J. S. 1989. The retention index and the rescaled consistency index. Cladistics 5:417–419.

Farris, J. S., V. A. Albert, M. Källersjö, D. Lipscomb, and A. G. Kluge. 1996. Parsimony jackknifing outperforms neighbor-joining. Cladistics 12:99–124.

Felsenstein, J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

Filichkin, S. A., and S. B. Gelvin. 1992. Effect of dimethyl sulfoxide concentration on specificity of primer matching in PCR. BioTechniques 126:828.

Hamby, R. K., and E. A. Zimmer. 1992. Ribosomal RNA as a phylogenetic tool in plant systematics. Pages 51–90 *in* Molecular systematics of plants II (P. S. Soltis, D. E. Soltis, and J. J. Doyle, eds.). Chapman and Hall, New York.

Hartmann, S., J. D. Nason, and D. Bhattacharya. 2001. Extensive ribosomal genic variation in the columnar cactus *Lophocereus*. J. Mol. Evol. 53:124–134.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.

Heenan, P. B., M. I. Dawson, and R. A. Bicknell. 2003. Evidence of apomictic seed formation in *Coprosma waina* (Rubiaceae). N.Z. J. Bot. 40:347–355.

Hershkovitz, M. A., E. A. Zimmer, and W. J. Hahn. 1999. Ribosomal DNA sequences and angiosperm systematics. Pages 268–326

in Molecular systematics and plant evolution (P. M. Hollingsworth, R. M. Bateman, and R. J. Gornall, eds.). Taylor & Francis, London.
Hillis, D. M., and M. T. Dixon. 1991. Ribosomal DNA: Molecular evolution and phylogenetic inference. Q. Rev. Biol. 26:411–453.
Huelsenbeck, J. P. 1995. The performance of phylogenetic methods in simulation. Syst. Biol. 44:17–48.
Huelsenbeck, J. P., B. Larget, R. E. Miller, and F. Ronquist. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst. Biol. 51:673–688.
Huelsenbeck, J. P., and F. Ronquist. 2001. MrBayes: Bayesian inference of phylogeny. Department of Biology, Univ. Rochester, Rochester, New York.
Imbert, F. M., and J. H. Richards. 1993. Protandry, incompatibility, and secondary pollen presentation in *Cephalanthus occidentalis* (Rubiaceae). Am. J. Bot. 80:395–404.
Karvonen, P., and V. O. Savolainen. 1993. Variation and inheritance of ribosomal DNA in *Pinus sylvestris* L.: Chromosomal organization and structure. Heredity 71:614–622.
Kiehn, M. 1985. Karyosystematische Untersuchungen an Rubiaceae: Chromosomenzahlung aus Afrika, Madagaskar und Mauritius. Plant Syst. Evol. 149:89–118.
Kiehn, M. 1986. Karyosystematic studies on Rubiaceae: Chromosome counts from Sri Lanka. Plant Syst. Evol. 154:213–223.
Kiehn, M. 1995. Chromosome survey of the Rubiaceae. Ann. Mo. Bot. Gard. 82:398–408.
Kiehn, M., and D. H. Lorence. 1996. Chromosome counts on angiosperms cultivated at the National Tropical Botanical Garden, Kaua'i, Hawaii. Pac. Sci. 50:317–323.
Kita, Y., and M. Ito. 2000. Nuclear ribosomal ITS sequences and phylogeny in East Asian *Aconitum* subgenus *Aconitum* (Ranunculaceae), with special reference to extensive polymorphism in individual plants. Plant Syst. Evol. 225:1–13.
Kluge, A. G., and J. S. Farris. 1969. Quantitative phyletics and the evolution of anurans. Syst. Zool. 18:1–32.
Kreader, C. A. 1996. Relief of amplification inhibition in PCR with bovine serum albumin or T4 gene 32 protein. Appl. Environ. Microbiol. 62:1102–1106.
Kuhner, M. K., and J. Felsenstein. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11:459–468.
Maluszynska, J., and J. S. Heslop-Harrison. 1991. Localization of tandemly repeated DNA on ribosomal DNA internal transcribed spacer (ITS) sequences in *Arabidopsis thaliana.* Plant J. 1:159–166.
Mayol, M., and M. Rossello. 2001. Why nuclear ribosomal DNA spacers (ITS) tell different stories in *Quercus.* Mol. Phylogenet. Evol. 19:167–176.
Mehra, P. N., and B. S. Gill. 1968. IOPB chromosome number reports XIX. Taxon 17:576.
Muir, G., C. C. Fleming, and C. Schlötterer. 2001. Three divergent rDNA clusters predate the species divergence in *Quercus petraea* (Matt.) Liebl. and *Quercus robur* L. Mol. Biol. Evol. 18:112–119.
Ohta, T., and G. A. Dover. 1983. Population genetics of multigene families that are dispersed into two or more chromosomes. Proc. Natl. Acad. Sci. USA 80:4079-4083.
Pop, M., and B. Oxelman. 2001. Inferring the history of the polyploid *Silene aegaea* (Caryophyllaceae) using plastid and homoeologous nuclear DNA sequences. Mol. Phylogenet. Evol. 20:474–481.
Posada, D., and K. A. Crandall. 1998. Modeltest: Testing the model of DNA substitution. Bioinformatics 14:817–818.
Puff, C., E. Robbrecht, R. Buchner, and P. De Block. 1996. A survey of secondary pollen presentation in the Rubiaceae. Opera Bot. Belg. 7:369–402.
Rambaut, A. 1996. Se-Al, version 1.dl. Sequence alignment program. Available at ftp://evolve.zo.ox.ac.uk/packages/Se-Al-All0a1.hqx.
Razafimandimbison, S. G., and B. Bremer. 2001. Tribal delimitation of Naucleeae (Rubiaceae): Inference from molecular and morphological data. Syst. Geogr. Plants 71:515–538.
Razafimandimbison, S. G., and B. Bremer. 2002. Phylogeny and classification of Naucleeae s.l. (Rubiaceae) inferred from molecular (ITS, *trn*T-L, *rbc*L) and morphological data. Am. J. Bot. 89:1027–1041.
Ridsdale, C. 1978. A revision of *Mitragyna* and *Uncaria* (Cinchoneae). Blumea 24:43–100.
Ritland, C. E., R. K. Ritland, and N. A. Straus. 1993. Variation in the ribosomal internal transcribed spacers (ITS1 and ITS2) among eight taxa of the *Mimulus guttatus* species complex. Mol. Biol. Evol. 10:1273–1288.
Rose, P. R., and B. T. Korber. 2000. Detecting hypermutations in viral sequences with an emphasis on G → A hypermutation. Bioinformatics 16:400–401.
Saghai-Maroof, K., M. Soliman, R. A. Jorgensen, and R. W. Allard. 1984. Ribosomal DNA spacer length polymorphism in barley: Mendelian inheritance, chromosomal location, and population dynamics. Proc. Natl. Acad. Sci. USA 81:8014–8018.
Sambrook, J., E. F. Fritsch, and T. Maniatis. 1989. Molecular cloning: A laboratory manual, 2nd edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
Sanderson, M. J., and J. J. Doyle. 1992. Reconstruction of organismal and gene phylogenies from data on multigene families: Concerted evolution, homoplasy, and confidence. Syst. Biol. 41:4–17.
Sang, T., D. G. 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 concerned evolution. Proc. Natl. Acad. Sci. USA 92:6813–6817.
Schlötterer, S., M.-T. Hauser, A. von Haeseler, and D. Tautz. 1994. Comparative evolutionary analysis of rDNA ITS regions in *Drosophila*. Mol. Biol. Evol. 11:513–522.
Schlötterer, S., and D. Tautz. 1994. Chromosomal homogeneity of *Drosophila* ribosomal DNA arrays suggests intrachromosomal exchanges drive concerted evolution. Curr. Biol. 4:777–783.
Shen, W. H., and B. Hohn. 1992. DMSO improves PCR amplification of DNA with complex secondary structure. Trends Genet. 87:227.
Smith, G. P. 1976. Evolution of repeated DNA sequences by unequal crossover. Science 191:528–535.
Suh, Y., L. B. Thien, H. E. Reeve, and E. A. Zimmer. 1993. Molecular evolution and phylogenetic implications of internal transcribed spacer sequences of ribosomal DNA in Winteraceae. Am. J. Bot. 80:1042–1055.
Swofford, D. L. 2000. PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4.0b. Computer program. Sinauer, Sunderland, Massachusetts.
Thompson, A. J. D., D. G. Higgins, and T. G. Gibson. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Computer program. Nucleic Acids Res. 22:4673–4680.
Urbatsch, L. E., B. G. Baldwin, and M. J. Donoghue. 2000. Phylogeny of the coneflowers and relatives (Heliantheae: Asteraceae) based on nuclear rDNA internal transcribed spacer (ITS) sequences and chloroplast DNA restriction site data. Syst. Bot. 25:539–565.
Wagner, A., N. Blackstone, P. Cartwright, M. Dick, B. Misof, P. Snow, G. P. Wagner, J. Batels, M. Murtha, and J. Pendleton. 1994. Surveys of gene families using polymerase chain reaction: PCR selection and PCR drift. Syst. Biol. 43:250–261.
Wendel, J. F., A. Schnabel, and T. Seelanan. 1995. Bidirectional interlocus concerted evolution following alloploid speciation in cotton (*Gossypium*). Proc. Natl. Acad. Sci. USA 92:280–284.
White, T. J., T. Bruns, S. Lee, and J. Taylor. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. Pages 315–322 *in* PCR protocols: A guide to methods and applications (M. A. Innis, D. H. Gelfand, J. J. Sninsky, and T. J. White, eds.). Elsevier, New York.
Yang, Y.-W., K.-N. Lai, P.-Y Tai, D.-P Ma, and W.-H. Li. 1999. Molecular phylogenetic studies of *Brassica*, *Rorippa*, *Arabidopsis* and allied genera based on the internal transcribed region of 18S-26S rDNA. Mol. Phylogenet Evol. 13:455–462.
Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol. 11:367–371.
Zimmer, E. A., S. L. Martin, S. M. Beverley, Y. W. Kan, and A. C. Wilson. 1980. Rapid duplication and loss of genes coding for the chains of hemoglobin. Proc. Natl. Acad. Sci. USA 77:2158–2162.
Zuker, M. 1989. On finding all suboptimal foldings of an RNA molecule. Science 244:48–52.

First submitted 3 December 2002; reviews returned 3 April 2003; final acceptance 5 December 2003
Associate Editor: François Lutzoni