J Appl Genet 51(3), 2010, pp. 259–273

Original article

Sequence characteristics and divergent evolution of the chloroplast *psbA-trnH* noncoding region in gymnosperms

D.C. Hao[1], S.L. Chen[2], P.G. Xiao[2]

[1]Laboratory of Biotechnology, Dalian Jiaotong University, Dalian, China
[2]Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences, Beijing, China

Abstract. The *psbA-trnH* intergenic region is among the most variable regions in the gymnosperm chloroplast genome. It is proposed as suitable for DNA barcoding studies and is useful in phylogenetics at the species level. This region consists of two parts differing in their evolutionary characteristics: 1) the *psbA* 3'UTR (untranslated region) and 2) the *psbA-trnH* intergenic spacer. We compared the sequence and RNA secondary structure of the *psbA* 3' UTR across gymnosperms and found consensus motifs corresponding to the stem portions of the RNA stem-loop structures and a consensus TGGATTGTTATGT box. The *psbA-trnH* spacer is highly variable in length and composition. Tandem repeats that form stem–loop structures were detected in both the *psbA* 3' UTR and the *psbA-trnH* spacer. The presence of promoters and stem–loop structures in the *psbA-trnH* spacer and high sequence variation in this region suggest that *psbA* and *trnH* in some gymnosperms are independently transcribed. A comparison of chloroplast UTRs across gymnosperms offer clues to the identity of putative regulatory elements and information on selective constraints imposed on the chloroplast non-coding regions. The present study should inspire researchers to explore the full potential of the *psbA-trnH* non-coding sequence and to further stimulate its application in a broader spectrum of studies, not limited to phylogenetics and DNA barcoding.

Keywords: DNA barcoding, *psbA-trnH* intergenic region, *psbA* 3' untranslated region, RNA secondary structure, stem-loop region

Introduction

Chloroplast DNA (cpDNA) sequence comparisons have been used widely as a tool in studies of plant phylogenetics and genome evolution. Among various cpDNA markers, the *psbA* (encodes photosystem II protein D1)-*trnH* (tRNA$^{\mathrm{His}}$) noncoding region is one of the most extensively used, particularly at the species level. This intergenic region consists of two evolutionarily distinct parts, i.e. the *psbA* 3'UTR, which is vital for posttranscriptional regulation of *psbA* gene expression, and the *psbA-trnH* intergenic spacer (IGS), which is highly variable. In recent years the psbA-trnH noncoding region has been employed as a candidate region for plant DNA barcoding. The *psbA-trnH* spacer, although short (approximately 450 bp), is the most variable plastid region in angiosperms and is easily amplified across a broad range of land plants (Kress et al. 2005). Kress et al. suggested that the sequences of *psbA-trnH,* along with nuclear ITS, have the potential to discriminate among the largest number of plant species for barcoding purposes. The *psbA-trnH* noncoding region was demonstrated to be successful as a DNA barcoding marker in angiosperms (Yao et al. 2009; Song et al. 2009) and now more extensive trials on non-flowering land plants, including gymnosperms, are required to verify its efficiency. Gymnosperms are unique in their evolutionary position and importance for conservation, and as such they need to be included

Received: August 8, 2009. Accepted: September 29, 2009.
Correspondence: D.C.Hao, Laboratory of Biotechnology, Dalian Jiaotong University, Dalian 116028, China; e-mail: haodc@126.com

in tests of proposed barcoding regions. It was found that neither the *psbA-trnH* region nor other proposed markers provided unique identifiers for all members of the Cycadales (Sass et al. 2007), and to date there has been no report studying the utility of the *psbA-trnH* region in other gymnosperms, which justifies a more in-depth investigation of sequence characteristics and evolution of the *psbA-trnH* region in gymnosperms. Many gymnosperm species are thought of as "living fossils" and the extant taxonomic assemblage represents only a sampling of the ancient diversity. Long-term evolution of gymnosperms might enable us to observe greater nucleotide divergence than one would expect in more recently derived species. Thus, the objectives of the current study are dual: 1) to study the spatial organization of the *psbA-trnH* intergenic region and quantify sequence divergence among and within phylogenetically diverse groups in gymnosperms, and 2) to test and evaluate the utility and limitation of the *psbA-trnH* intergenic region for gymnosperm DNA barcoding.

Materials and methods

DNA sequences and alignments

Our dataset was derived from the complete set (113) of the gymnosperm *psbA-trnH* noncoding sequences. Sequences of Gnetales, Welwitschiales, Ephedrales, Cupressaceae, Pinaceae, Araucariaceae, and Cycadales were retrieved from the NCBI GenBank. Accession numbers used in this study are listed in table S1. Genomic DNA of Taxaceae (28), Cephalotaxaceae (14), and Podocarpaceae (2) species was extracted using a Universal Genomic DNA Extraction Kit (Takara, Dalian, China). A 50 µL PCR reaction mix consisted of 5 µL of 10× reaction buffer, 4 µL each 2.5 mM dNTP stock, 2.5 µL of 10 µM forward and reverse primers, and 1.5 U Ex Taq polymerase (Takara, Dalian, China). Approximately 50 ng genomic DNA were used as a template for the reaction. The reaction mixture was placed in a Takara PCR Thermal Cycler Dice (Takara, Japan). The primers used for amplification of *psbA-trnH* (psbA3'f: 5'-GTTATGCATGAAC GTAATGCTC and trnHf: 5'-CGCGCATGGTG GATTCACAATCC) and the cycling (38 cycles) conditions were described previously (Kress et al. 2005). DNAs were purified using an Agarose Gel DNA Purification Kit (Takara).

All PCR products were subcloned into a TA cloning vector pMD19-T (Takara). The plasmids were purified for sequencing. An ABI Prism, a BigDye Terminator, and a Cycle Sequencing Ready Reaction Kit (Applied Biosystems, Foster City, CA) were used for the sequencing reaction with RV-M and M13-47 primers. The sequences were detected using an ABI Prism 377 Genetic Analyzer (Applied Biosystems). The obtained sequences were aligned and edited using Muscle (Edgar 2004; http://www.drive5.com/muscle/), Clustal W2, and Bioedit (Hall 1999). WebLogo 3 (Crooks et al. 2004) was used to visualize conserved regions from the gymnosperm-level multiple alignment. Lengths and A+T contents were calculated from the aligned sequences. Sequence divergences were calculated using the maximum composite likelihood (MCL) model in MEGA4 (Tamura et al. 2007).

RNA structure, repeat sequence analyses, promoter prediction

RNAfold in the Vienna package version 1.6.1 and mfold version 3.2 (http://frontend.bioinfo.rpi.edu/) were used to predict structures using the default parameters. RNAfold was also used to measure the minimal free energy (MFE) for each sequence with its default parameters. It predicts the free energy of the most stable RNA structure for a given sequence. The base-pair probabilities were calculated by RNAfold as well (McCaskill 1990). RNAz (Gruber et al. 2007) was used to detect thermodynamically stable and evolutionarily conserved RNA secondary structures in multiple sequence alignments. Inverted repeats were found with einverted (http://mobyle.pasteur.fr/cgi-bin/portal.py?form=einverted). The Tandem Repeats Finder program (http://tandem.bu.edu/trf/trf.html; Benson 1999) was used to detect direct repeats. To detect promoter elements of the tRNA genes, we used Neural Network Promoter Prediction (http://www.fruitfly.org/seq_tools/promoter.html) to examine the IGS sequences for ‘‘–35’’ and ‘‘–10’’ prokaryotic promoter element homologies and TSSP-TCM (http://mendel.cs.rhul.ac.uk/ mendel.php?topic=fgen) to search for promoters in plant sequences.

Phylogenetic analysis

The best-fit evolutionary model and the gamma shape parameter of among-site rate variation were inferred with ModelTest 3.8 (Posada 2006); the latter was used to calculate the transition/transversion ratio (R) with MEGA4. Distances were estimated using the pairwise-deletion option and standard errors were calculated by the bootstrap method with 1,000 replicates. The presence of selection was tested in *psbA-trnH* regions using Tajima's neutrality test statistic D under the H_0 hypothesis: neutral mutation, no selection, being an alternative H_1 hypothesis: the presence of selection (Tajima 1989; Tamura et al. 2007).

We designated *Podocarpus* as a functional outgroup for the phylogenetic analysis of Taxaceae and Cephalotaxaceae (Hao et al. 2008, 2009). We used MEGA4, GARLI (Zwickl 2006), and MrBayes 3.1.2 (Ronquist and Huelsenbeck 2003) for phylogenetic analyses. The data matrix of the *psbA-trnH* spacer was analyzed by neighbor-joining (NJ), maximum likelihood (ML), and Bayesian inference. NJ used maximum composite likelihood (MCL) distances and pairwise deletion of gaps. ML searches relied on the respective evolutionary model (Table 1) of each gymnosperm group, which ModelTest selected as the best-fitting model. For example, HKY+G was used for Cupressaceae and GTR for Cephalotaxaceae, respectively. Bayesian probabilities were obtained with four Markov chain Monte Carlo chains run for 700 thousand generations, using random trees as the starting point, and sampling every 100th generation. The trees sampled before the saturation of maximum likelihood estimates were discarded as burn-in. Nonparametric bootstrap support for ML and NJ was obtained by resampling the data 1,000 times with the same search options and model.

Results and discussion

Length, AT content and other sequence characteristics

The length of the *psbA-trnH* region (start from stop codon TAG in Cycadales or TAA in the other groups) was 573±126.2 bp for Taxaceae, 268±0.27 bp for Cephalotaxaceae, 278±12.7 bp for Cycadales, 463±49.1 bp for Cupressaceae, 474 ±4.8 bp for Ephedrales, 516±40.7 bp for Pinaceae, 562±6.1 bp for Gnetales, and 807±246.1 bp for Podocarpaceae+Araucariaceae (Figure 1A). The *psbA-trn H* region of Taxaceae is longer than that of Cephalotaxaceae and shorter than that of Podocarpaceae+Araucariaceae (one-way ANOVA: F=20.93, P<0.0001; Tukey HSD-test for pairwise comparison, P<0.01). The length of the *psbA-trnH* region in Taxaceae, Ephedrales, and Pinaceae was not significantly different (P>0.05). Within Taxaceae, there is no significant length difference among *Taxus, Torreya,* and *Amentotaxus* (one-way ANOVA: F=2.97, P=0.069). Length variation in gymnosperms mainly results from multiple insertions/deletions in the *psbA-trnH* intergenic spacer, while the length of *psbA* 3'UTR remains constant (< 200 bp, see below).

The highest A+T content of the *psbA-trnH* region was 67.8%±2.38% for Taxaceae and the lowest was 61.6%±1.4% for Gnetales (Figure 1B). The A+T content was not significantly different among eight gymnosperm groups (one-way ANOVA: F =1.92, P=0.115). Within Taxaceae, although the A+T content of the *psbA-trnH* region was higher in *Taxus* (69.3%±1.88%) than in the other genera, the difference is not statistically sig-

Table 1. Sequence characteristics of gymnosperm *psbA-trnH* regions

	Gnetales+ Welwitschiales	Ephedrales	Cupressaceae	Taxaceae	Cephalo -taxaceae	Pinaceae	Podocarpaceae +Araucariaceae	Cycadales
Overall average distance[a]	0.088 ±0.014	0.001 ±0	0.109 ±0.012	0.136 ±0.016	0.016 ±0.005	0.107 ±0.01	0.145 ±0.52	0.008 ±0.004
R[b]	0.894	176.8	0.798	0.766	9.92	1.075	2.822	0
Transition/transversion rate ratios	k1[c]=1.606 k2[d]=2.962	k1=1 k2=1000	k1=2.15 k2=2.493	k1=1.986 k2=2.565	k1=48.057 k2=3.5	k1=2.578 k2=2.816	k1=10.325 k2=5.033	k1=0 k2=0
π[e]	0.0815	0.0008	0.0703	0.0991	0.0134	0.0891	–	–
D[f]	2138853	-1.457	0.0726	0.425	-0.971	0.335	–	–
Evolutionary model	K81uf	TrN	HKY+G	K81uf+G	GTR	TVM+I	K81uf+G	K81uf

[a] mean ± SE of overall average genetic distance calculated by MEGA4

[b] overall transition/transversion bias

[c] k1 (purines)

[d] k2 (pyrimidines)

[e] nucleotide diversity

[f] Tajima test statistic

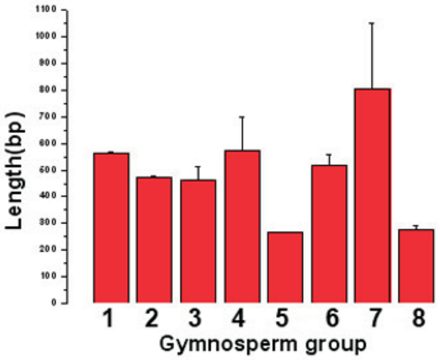


Figure 1A. Sequence and structure characteristics of the gymnosperm *psbA-trnH* intergenic regions. Length variation. Bar represents standard deviation of the average. 1, Gnetales+Welwitschiales; 2, Ephedrales; 3, Cupressaceae; 4, Taxaceae; 5, Cephalotaxaceae; 6, Pinaceae; 7, Podocarpaceae +Araucariaceae; 8, Cycadales

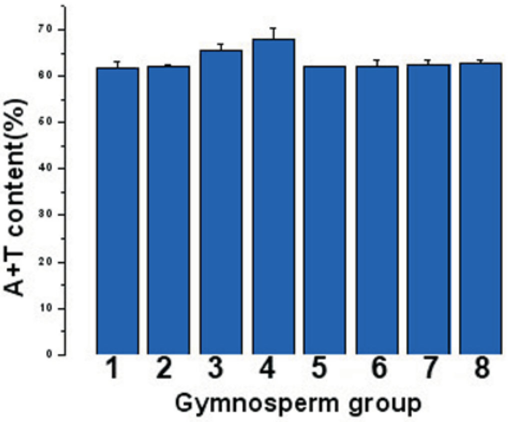


Figure 1B. Sequence and structure characteristics of the gymnosperm *psbA-trnH* intergenic regions. Bar represents standard deviation of the average. 1, Gnetales+Welwitschiales; 2, Ephedrales; 3, Cupressaceae; 4, Taxaceae; 5, Cephalotaxaceae; 6, Pinaceae; 7, Podocarpaceae +Araucariaceae; 8, Cycadale A+T content variation,

nificant (one-way ANOVA: F=1.14, P=0.335). The AT-rich region might contain a promoter sequence and other regulatory elements and we also wondered whether *trnH* is transcriptionally independent; thus we used multiple programs to detect a potential promoter region containing a putative transcription initiation site. Potential promoter sequences were not found in the spacer regions of Gnetales, Welwitschiales, Ephedrales, and Cephalotaxaceae. In contrast, conserved elements (AAGGAAATA) with high similarity to bacterial sigma^{70}-type promoters were detected in Araucariaceae (Table S2) using a support vector machine (SVM) approach (SAK; Gordon et al. 2003). Plant RNA polymerase II promoters were also found in Araucariaceae with the use of the SVM approach (tsspTCM; Shahmuradov et al. 2005). The time-delay neural network approach (BDGP; Reese 2001) found eukaryotic promoters in Cupressaceae, Taxaceae, Pinaceae, Podocarpaceae, Araucariaceae and Cycadaceae. Secondary structure calculations revealed that these promoter elements are involved in forming the stem and/or loop of the stable stem-loop structures (data not shown). Compared to Taxaceae, Pinaceae, and Cycadaceae, secondary structures of *Araucaria* have the lowest MFE (–283.1±59.2 kcal mol^{-1}, Figure 1C; one-way ANOVA: F=20.72, P<0.0001; Tukey HSD-test for pairwise comparison, P<0.01) and thus they might be more stable. The *psbA-trnH* spacer of these gymnosperms can be regarded as the starting point for the transcription of the tRNA$^{\text{His}}$; however, functionality of promoters needs to be proven by experimental data.

The Tandem Repeats Finder program detected 8, 3, 1, and 1 putative repeats in Taxaceae, *Araucaria*, *Cycas,* and Cupressaceae (Table S3), respectively. The repeat sequences were 13–95 bp in length, had 1.9–4.0 copies, and their match points were between 0.96 and 1.0. It is worthy of note that the repeat sequences of *Taxus* are significantly longer than those of the other gymnosperms

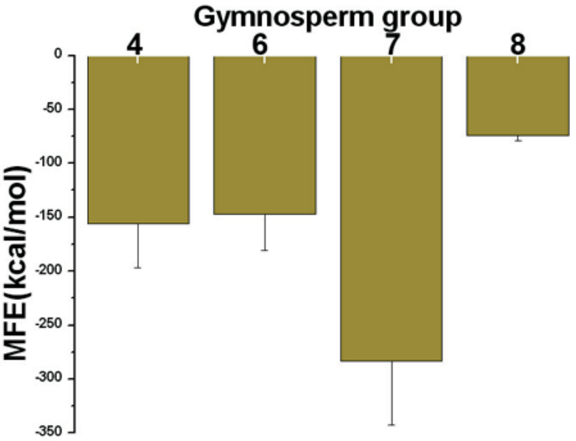


Figure 1C. Sequence and structure characteristics of the gymnosperm *psbA-trnH* intergenic regions. Minimal free energy of secondary structures. Bar represents standard deviation of the average. 1, Gnetales+Welwitschiales; 2, Ephedrales; 3, Cupressaceae; 4, Taxaceae; 5, Cephalotaxaceae; 6, Pinaceae; 7, Podocarpaceae +Araucariaceae; 8, Cycadales

(one-way ANOVA: F=35.29, P<0.0001; Tukey HSD-test for pairwise comparison, P<0.01). The prevalence of repeat sequences in the Taxaceae *psbA-trnH* spacer is reminiscent of large amounts of repeat sequences detected in the Taxaceae *trnL-F* spacer (Hao et al. 2009), implying a similar mechanism, i.e. slipped-strand mispairing. Interestingly, closely related species have identical or similar repeat sequences, e.g. both *T. yunnanensis* and *T. wallichiana* have a repeat sequence of 69 bp, and the former has one more copy than the latter (Table S3); *T. cuspidata* and two related hybrid species, *T.* × *media* and *T.* × *hunnewelliana*, have a repeat sequence of 95 bp, and the latter two have one more copy than the former; *T. canadensis* has a repeat sequence of 93 bp that lacks one "AT" compared to the repeat sequence of *T. cuspidata*. These observations are consistent with the generally accepted view that tandem arrays of perfect repeats are hotspots for replication errors, resulting in high rates of expansions/contractions. Whether expansions/contractions in the number of perfect repeat units in the *trnH* promoter are associated with variable transcription of the *trnH* needs to be investigated further in more details.

General feature of the *psbA* 3' UTR in gymnosperms

A detailed inspection of a logo (Figure 2) suggests high conservancy across nearly an entire UTR. Two regions of similarity across all gymnosperm *psbA 3*' UTRs were identified (Figure 2). The first was between the stop codon of *psbA* and a 3' UTR stem-loop structure. A short motif TGGATTGT TATGT was conserved across gymnosperms (Figure 2), which is longer than and different from the conserved motif found in angiosperms (Storchova and Olson 2007). Conservation of this sequence motif may reflect its functional importance, although it is unknown how deletion of this sequence motif influences mRNA longevity and transcript processing in gymnosperms. The second region of similarity was associated with the stem portion of the predicted RNA stem-loop structure (Figure 2 and Table 2). The stem could be defined by two consensus sequence motifs: 1) AGTACCAA ("1" in Figure 2) and complementary motif TTGGTACT ("1'"), located in the upper part of the stem, and 2) AAGAAAAAAA ("2" in Figure 2) and complementary TTTTTTTCTT ("2'") found in the lower stem (Figure 2). Bollenbach and Stern (2003) found that secondary structures common to chloroplast mRNA 3'-untranslated regions direct cleavage of stem-loop-containing RNAs by CSP41, an endoribonuclease belonging to the short chain dehydrogenase/reductase superfamily. The pattern of conservation of the 3' end of *psbA* that forms the stem-loop is consistent with the importance of this structure for mRNA stability. In contrast to the high levels of similarity across gymnosperms in the stem-forming regions, the loop and bulge regions were highly variable, sug-

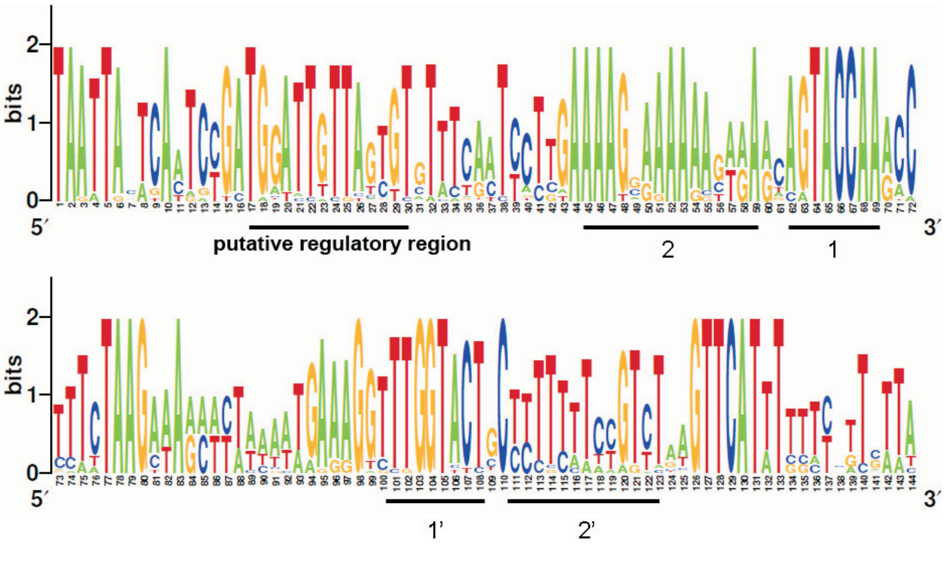
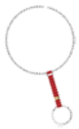
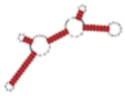
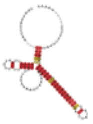
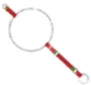
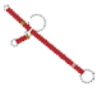
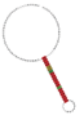
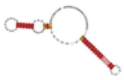
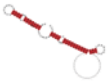


Figure 2. Consensus motifs in 3' UTR of *psbA.* This sequence logo was generated from the multiple alignment of sequences from 26 genera representing all gymnosperm families (10) with *psbA-trnH* sequence records in the GenBank. The logo displays the consensus sequence, the relative frequency of nucleotides and information content (measured in bits) at every position of sequence. If a specific nucleotide is present at the respective position in 100% of accessions, information content is equal to two. The logo starts with the TAA stop codon of the *psbA*. The putative regulatory region and sequences involved in the formation of stem-loop secondary structures are indicated by dark bars.

Table 2. RNAz analysis of secondary structures in 3’ UTR of gymnosperm *psbA* mRNA

	Gnetales + Welwitschiales	Ephedrales	Cupressaceae	Taxaceae	Cephalotaxaceae	Pinaceae	Podocarpaceae +Araucariaceae	Cycadales
Location	0–120	0–120	0–120	0–120	0–120	0–120	0–120	0–120
Mean pairwise identity	74.64	100	75.33	79.45	94.11	80	81.59	81.98
Mean single sequence MFE	−22.35	−30.70	−34.45	−31.28	−45.55	−40.52	−37.82	−27.17
Consensus MFE	−15.49	−30.70	−15.98	−20.70	−45.08	−30.19	−25.95	−16.82
Structure conservation index	0.69	1	0.46	0.66	0.99	0.75	0.69	0.62
Secondary Structure								

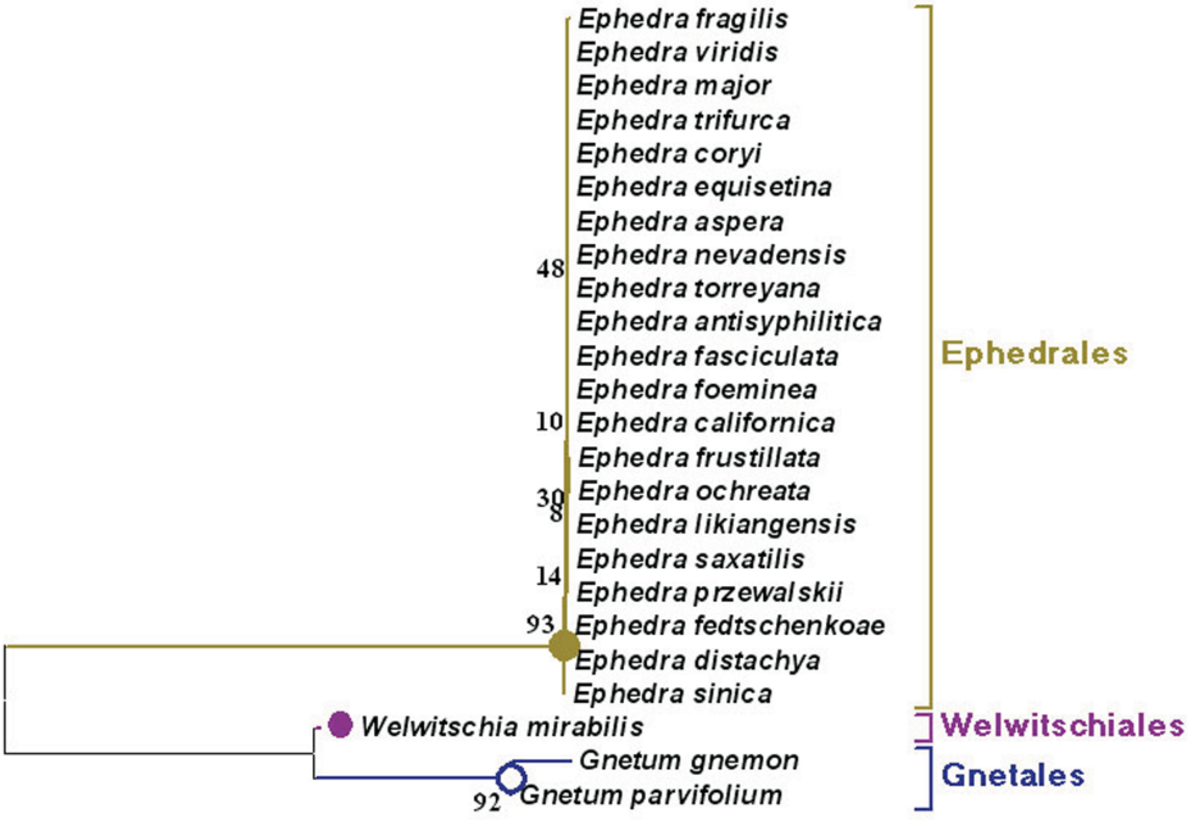


Figure 3A. Phylogenetic relationship of gymnosperm *psbA-trnH* intergenic regions revealed by NJ trees. Numbers beside branches are bootstrap values. Gnetales, Welwitschiales, and Ephedrales

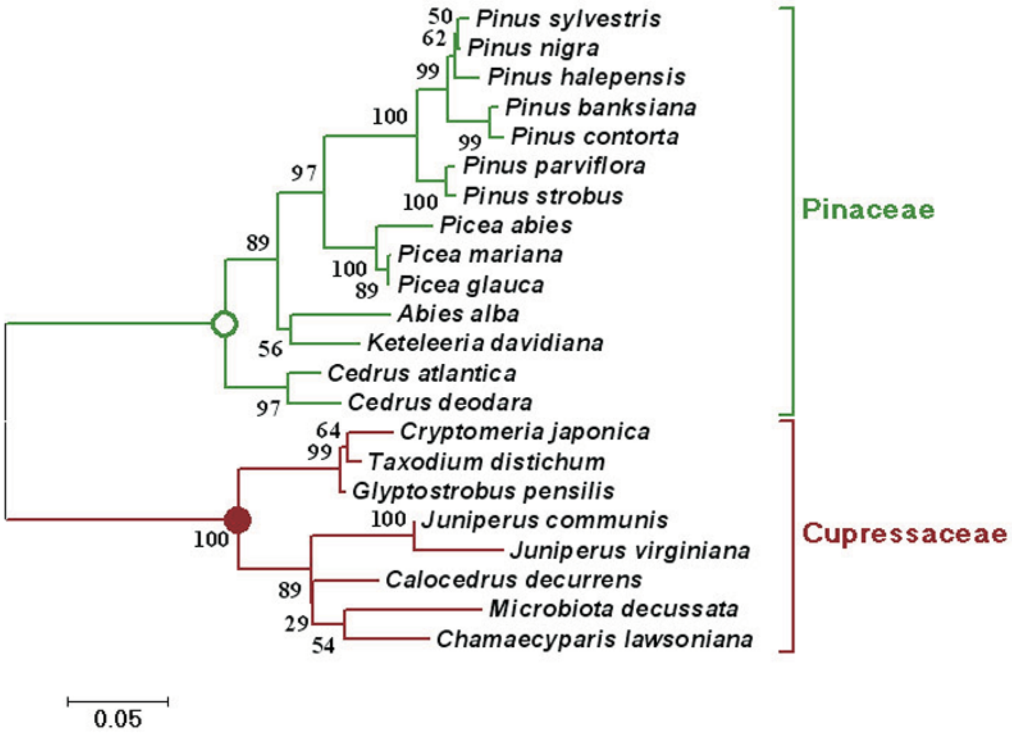


Figure 3B. Phylogenetic relationship of gymnosperm *psbA-trnH* intergenic regions revealed by NJ trees. Numbers beside branches are bootstrap values. Cupressaceae and Pinaceae

gesting that they are less functionally constrained (Figure 2 and data not shown).

Phylogenetic tree and the utility of the psbA-trnH region in DNA barcoding

The strategy of using limited DNA information to aid species identification, i.e. DNA barcoding, is relatively attractive for many practical, commercial and scientific applications. The *psbA-trnH* intergenic spacer is one of two earliest proposed markers (Kress et al. 2005). The *psbA-trnH* intergenic spacer is also the chloroplast marker used most extensively for DNA barcoding (Song et al. 2009; Yao et al. 2009). Therefore, this marker was tested in this study for its utility in generating unique identifiers for gymnosperms. Sequences from PCR and sequencing (28 Taxaceae, 14 Cephalotaxaceae, and 2 Podocarpaceae taxa) as well as those acquired from the GenBank were subjected to phylogenetic analyses, and a NJ tree generated by MEGA4 is shown in Figure 3A–D. Bayesian analysis and the ML method generated virtually the same topology as that shown in Figure S1. Taxaceae and Cephalotaxaceae are sister clades, if Podocarpaceae are regarded as an outgroup. Within Cephalotaxaceae, the phylogenetic relationship was not resolved, except that *C. griffithii* and *C. oliveri* form a basal group and *C. latifolia* is between this group and the other *Cephalotaxus*. Within Taxaceae there are two sister clades: one consisting of *Torreya* and *Amentotaxus*, and the other consisting of *Taxus, Pseudotaxus,* and *Austrotaxus*. The relationship within the respective genera is well resolved and does not contradict previous studies with multiple molecular markers (Hao et al. 2008), except that *psbA-trnH* of *Taxus floridana* is closer to that of *Torreya grandis* than to other Taxus (Figure 3C). The topology of the *psbA-trnH* tree may reflect the suitability of this marker for DNA barcoding in Taxaceae, but not in Cephalotaxaceae. Accordingly, the average genetic distance of *psbA-trnH* in different taxa within Taxaceae is 0.136±0.016, which is much larger than that within Cephalotaxaceae (0.016±0.005, Table 1). Similarly, *psbA-trnH* could not be a candidate marker of DNA barcoding for Ephedrales (Figure 3A) and Araucariaceae (Figure 3D), whereas it could be suitable for Cupressaceae and Pinaceae (Figure 3B). Whether it is valid for Gnetales and Podocarpaceae is worth further study. The *psbA-trnH* spacer primers specified by Kress et al. (2005) yielded distinct double bands in all

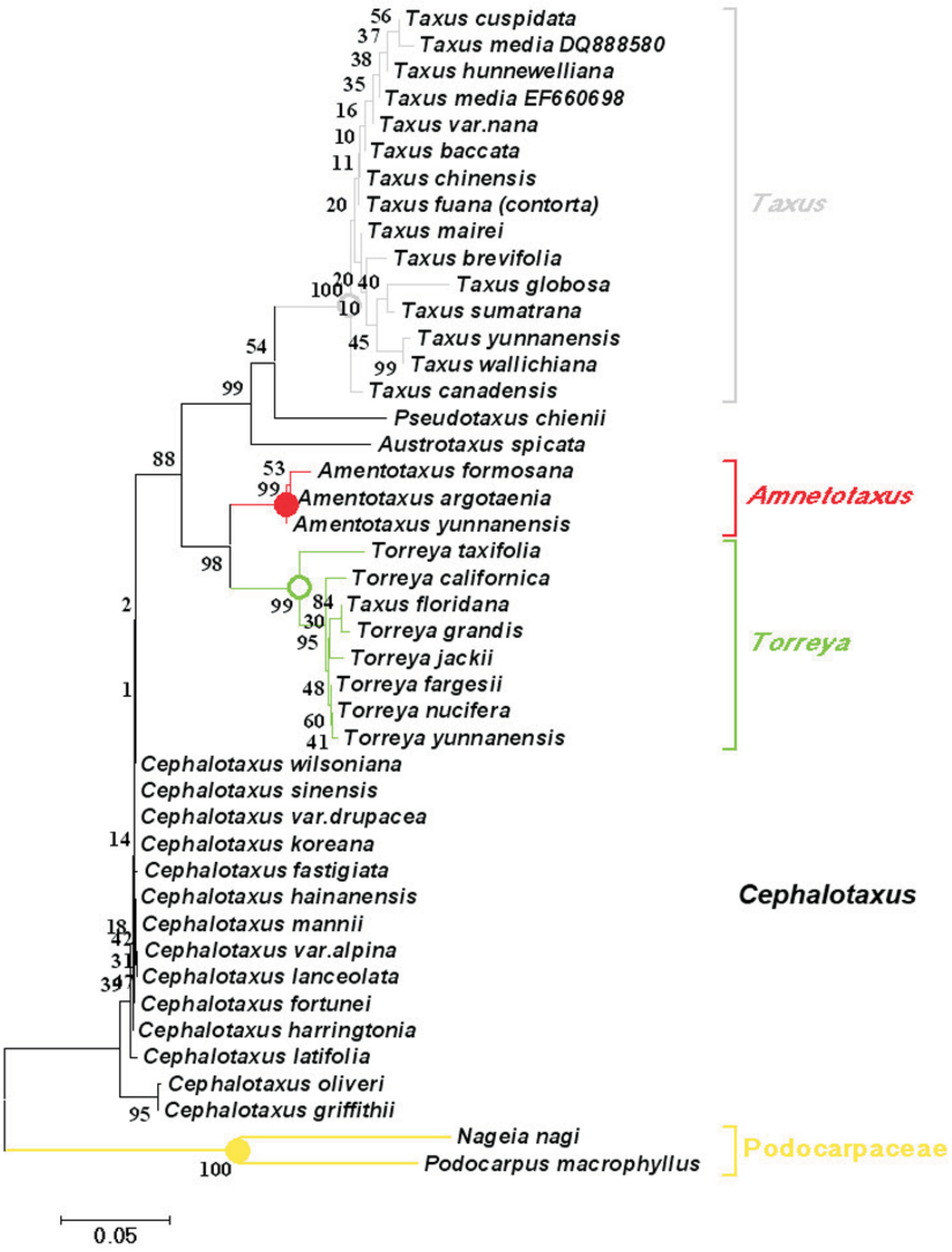


Figure 3C. Phylogenetic relationship of gymnosperm *psbA-trnH* intergenic regions revealed by NJ trees. Numbers beside branches are bootstrap values. Taxaceae, Cephalotaxaceae, and Podocarpaceae

Cycadales species but *Cycas* (Sass et al. 2007), and the addition of *psbA-trnH* sequence data did not further resolve the non-specific identification made by nrITS for the species tested. This is in accordance with our finding that *psbA-trnH* might not be used as the barcoding marker of Cycadales (Figure 3D).

In conclusion, complexity of evolutionary patterns in non-coding sequences, such as *psbA-trn*H non-coding sequences, is largely caused by frequent micro-structural mutations in addition to substitutions of nucleotides. A significant sequence divergence makes it fruitful to use *psbA-trnH* in gymnosperm DNA barcoding studies, especially for groups such as Cupressaceae, Taxaceae and Pinaceae. With more *psbA-trnH* sequences in public databases, we would be able to have a thorough examination of its utility in more

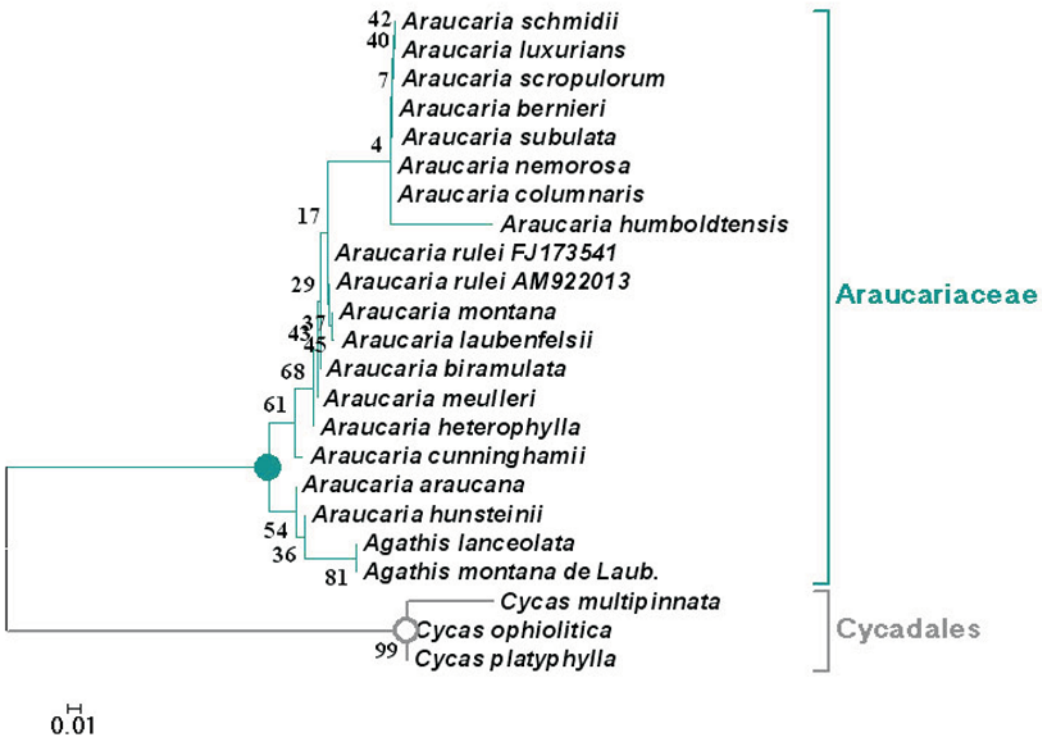


Figure 3D. Phylogenetic relationship of gymnosperm *psbA-trnH* intergenic regions revealed by NJ trees. Numbers beside branches are bootstrap values. Araucariaceae and Cycadales

gymnosperm groups. In spite of the divergent evolution of the *psbA-trnH* non-coding sequence, there is a consensus secondary stem-loop structure in the 3' UTR of *psbA*, implying purifying selection. The present study should inspire researchers to explore the full potential of the *psbA-trnH* non-coding sequence and further stimulate its application in a broader spectrum of studies, not limited to phylogenetics and DNA barcoding.

Acknowledgements. This study was sponsored by the Education Department of the Liaoning Province (2009A120), and a Start-up research fund (2008-2010) of the Dalian Jiaotong University.

REFERENCES

Bollenbach TJ, Stern DB, 2003. Secondary structures common to chloroplast mRNA 3'-untranslated regions direct cleavage by CSP41, an endoribonuclease belonging to the short chain dehydrogenase/reductase superfamily. J Biol Chem 278: 25832–25838.

Crooks GE, Hon G, Chandonia JM, Brenner SE, 2004. WebLogo: A sequence logo generator. Genome Res 14: 1188–1190.

Edgar RC, 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl Acids Res 32: 1792–1797.

Gordon L, Chervonenkis AY, Gammerman AJ, Shahmuradov IA, Solovyev VV, 2003. Sequence alignment kernel for recognition of promoter regions. Bioinformatics 19: 1964–1971.

Gruber AR, Neuböck R, Hofacker IL, Washietl S, 2007. The RNAz web server: prediction of thermodynamically stable and evolutionarily conserved RNA structures. Nucl Acids Res 35: W335–338.

Hall TA, 1999. BIOEDIT: a user-friendly biological sequence alignment, editor and analysis program for Windows 95/98/NT. Nucl Acids Symposium Series 41: 95–98.

Hao DC, Huang B, Yang L, 2008. Phylogenetic relationships of the genus *Taxus* inferred from chloroplast intergenic spacer and nuclear coding DNA. Biol Pharm Bull 31: 260–265.

Hao DC, Huang B, Chen SL, Mu J, 2009. Evolution of the chloroplast *trnL-trnF* region in the gymnosperm lineages Taxaceae and Cephalotaxaceae. Biochem Genet 47: 351–369.

Hao DC, Xiao PG, Huang B, Ge GB, Yang L, 2008. Interspecific relationships and origins of Taxaceae and Cephalotaxaceae revealed by partitioned Bayesian analyses of chloroplast and nuclear DNA sequences. Plant Syst Evol 276: 89–104.

Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH, 2005. Use of DNA barcodes to identify flowering plants. Proc Natl Acad Sci USA 102: 8369–8374.

McCaskill JS, 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers 29: 1105–1119.

Posada D, 2006. ModelTest Server: a web-based tool for the statistical selection of models of nucleotide substitution online. Nucl Acids Res 34: W700–W703.

Reese MG, 2001. Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome. Comput Chem 26: 51–56.

Ronquist F, Huelsenbeck JP, 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572–1574.

Sass C, Little DP, Stevenson DW, Specht CD, 2007. DNA barcoding in the cycadales: testing the potential of proposed barcoding markers for species identification of cycads. PLoS One 2: e1154.

Shahmuradov IA, Solovyev VV, Gammerman AJ, 2005. Plant promoter prediction with confidence estimation. Nucl Acids Res 33: 1069–1076.

Song J, Yao H, Li Y, Li X, Lin Y, Liu C, et al. 2009. Authentication of the family Polygonaceae in Chinese pharmacopoeia by DNA barcoding technique. J Ethnopharmacol 124: 434–439.

Storchova H, Olson MS, 2007. The architecture of the chloroplast *psbA-trnH* non-coding region in angiosperms. Pl Syst Evol 268: 235–256.

Tajima F, 1989. Statistical methods to test for nucleotide mutation hypothesis by DNA polymorphism. Genetics 123: 585–595.

Tamura K, Dudley J, Nei M, Kumar S, 2007. Mega4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol Biol Evol 24: 1596–1599.

Yao H, Song JY, Ma XY, Liu C, Li Y, Xu HX, et al. 2009. Identification of *Dendrobium* species by a candidate DNA barcode sequence: The chloroplast *psbA-trnH* intergenic region. Planta Med 75: 667–669.

Zuker M, 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucl Acids Res 31: 3406–3415.

Zwickl DJ, 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. dissertation, University of Texas, Austin.

Supplementary material

Table S1

Sampling design

Group No	Group	Order	Family	Genus	Species	GenBank No
1	2	3	4	5	6	7
1	Gnetales	Gnetales	Gnetaceae	Gnetum	gnemon	AY849369
					parvifolium	NC_011942
		Welwitschiales	Welwitschiaceae	Welwitschia	mirabilis	AY849370
2	Ephedrales	Ephedrales	Ephedraceae	Ephedra	antisyphilitica	AY849359
					aspera	AY849365
					californica	AY849358
					equisetina	AY849352
					coryi	AY849366
					foeminea	AY849353
					fasciculata	AY849360
					fragilis	AY849363
					frustillata	AY849355
					distachya	AY849351
					fedtschenkoae	AY849350
					likiangensis	AY849357
					major	AY849361
					nevadensis	AY849354
					ochreata	AY849362
					przewalskii	AY849348
					sinica	AY849349
					saxatilis	AY849364
					torreyana	AY849356
					trifurca	AY849368
					viridis	AY849367
3	Coniferales-1	Coniferales	Cupressaceae	Juniperus	communis	EU750613-EU750616
					virginiana	EU750617-EU750620
				Calocedrus	decurrens	FJ493277
				Chamaecyparis	lawsoniana	FJ493278
				Microbiota	decussata	AM887665
						AM887666
						FM205067-FM205112
				Cryptomeria	japonica	AY727189
				Glyptostrobus	pensilis	AY727190
				Taxodium	distichum	AY727188
4	Coniferales-2	Coniferales	Cephalotaxaceae	Cephalotaxus	harringtonia	**EF660677**
					wilsoniana	**EF660674**
					sinensis	**EF660687**
					fortunei	**EF660695**
					latifolia	**EF660686**
					lanceolata	**EF660676**
					hainanensis	**EF660688**
					oliveri	**EF660701**
					var.alpina	**EF660680**
					mannii	**EF660675**
					griffithii	**EF660669**
					var.drupacea	**EF660684**
					koreana	**EF660703**
					fastigiata	**EF660689**
			Taxaceae	Amentotaxus	argotaenia	**EF660691**
					formosana	**EF660670**

Table S1 cont.

1	2	3	4	5	6	7
					yunnanensis	**EF660681**
				Austrotaxus	spicata	**EF660671**
				Pseudotaxus	chienii	**EF660683**
				Taxus	baccata	**EF017303**
					brevifolia	**EU078560**
					mairei	**DQ888577**
					cuspidata	**DQ888579**
					var.nana	**EF660682**
					yunnanensis	**EF660668**
					chinensis	**DQ888576**
					×hunnewelliana	**EF017302**
					wallichiana	**EF660700**
					fuana (contorta)	**EF660685**
					sumatrana	**EF660672**
					×media	**DQ888580**
						EF660698
					canadensis	**EF017304**
					floridana	**EF660679**
					globosa	**EF660673**
				Torreya	yunnanensis	**EF660678**
					nucifera	**EF660697**
					taxifolia	**EF660702**
					fargesii	**EF660694**
					californica	**EF660699**
					grandis	**EF660692**
					jackii	**EF660693**
5	Coniferales-3	Coniferales	Pinaceae	Picea	abies	FJ493294
					mariana	EU750626
					glauca	EU750621-EU750624
				Abies	alba	FJ493291
				Cedrus	atlantica	FJ493292
					deodara	FJ493293
				Keteleeria	davidiana	NC_011930
				Pinus	sylvestris	FJ493296
					banksiana	EU750628
					halepensis	EU531714
					contorta	X57097
					nigra	FJ493295
					parviflora	EF590724
					strobus	EU750631
6	Coniferales-4	Coniferales	Podocarpaceae	Nageia	nagi	**EF660696**
				Podocarpus	macrophyllus	**EF660690**
7	Coniferales-5	Coniferales	Araucariaceae	Agathis	lanceolata	AM921997
					montana de Laub.	AM921998
				Araucaria	araucana	AM922001
					bernieri	AM922002
					biramulata	FJ173522
					columnaris	AM922005
					cunninghamii	AM922006
					heterophylla	FJ173525
					humboldtensis	FJ173527
					hunsteinii	FJ173528
					laubenfelsii	FJ173530
					luxurians	FJ173533
					meulleri	AM922010
					montana	FJ173537
					nemorosa	AM922011

Table S1 cont.

1	2	3	4	5	6	7
					rulei	AM922013
						FJ173541
					scropulorum	FJ173548
					schmidii	FJ173545
					subulata	AM922014
8	Cycadales	Cycadales	Cycadaceae	Cycas	multipinnata	EF612963
					ophiolitica	EF612962
					platyphylla	EF612961

New psbA-trnH sequences from this study are in bold type

Table S2. Predicted promoter sequences in the *psbA-trnH* noncoding region of gymnosperms

Taxon	Start-end (nucleotide position)	Score	Promoter sequence	Prediction program
1	2	3	4	5
Cupressaceae				
Juniperus virginiana	28-78	0.90	ATTTCAATCCTAAAAAAGCAGTACCAATTTGGTACTGCTTTTTCCGTCTA	BDGP
Taxaceae				
Austrotaxus spicata	221-271	0.81	GATAAAGCAATAAAAAAGTTGCTACTACTTAGAGATTAAGTAGCAACTTA	BDGP
Pseudotaxus chienii	35-85	0.86	CAATCCTGTAAAAAAAAGTACCAAGCCTTTCAAAATCAAAAAGGCTTGGT	BDGP
Taxus yunnanensis	375-425	0.89	TATACTTTTATATAAAATGATGACAATTAGACTATAAATAGATATAATCT	BDGP
Taxus wallichiana	374-424			
Taxus sumatrana	247-297	0.91	AATATATCTATATATATTACCTTATATTAGGTACCCAATCTGATTCTCTT	BDGP
Taxus globosa	244-294	0.98	ATATAATATATATATATATATTACCTTATATTAGGTACCCAATCTGATTC	BDGP
Pinaceae				
Cedrus atlantica	113-163	0.90	TTCCCATTCTATAAAGAATGGATATGTGCAGTTCCCCTGCATCCAGCAGG	BDGP
Keteleeria davidiana	475-525	0.96	TTTTTTTTTGTAAAAAAGAACCGTGGACCGTGGATAGAGACAATTGGTTT	BDGP
Pinus banksiana	165-215	0.99	GACTCAGATCTAAAATTGGGCGGGATTGGGACCCATTTATATTCTTTCTC	BDGP
Pinus contorta	254-304			
Pinus strobus	369-419	0.92	ATTTCATTTTTATAATAAGCCGAACAACTTGTTCGAGAGTTGGGAGTTAG	BDGP
Podocarpaceae				
Podocarpus macrophyllus	227-277	0.97	CCCCCGATCTGTATATACCCTCTGCGCTGAAGGAAAGCGCACAGATATAG	BDGP
Araucariaceae				
Araucaria araucana	154-204	0.93	CCATCTGGACTATAAACCCAGATGGTAAATCCGTCCGTCCAATTGAGACT	BDGP
	435-459	0.97	TATATATATCTATTACCTATAGATA	tsspTCM
	479-489	1.47	TCAAGGAAATA	SAK
Araucaria bernieri	516-566	0.99	CCCCGATATATATATATATAGATAGAGATATATATATATATCTATTACCT	BDGP/tsspTCM
Araucaria biramulata/columnaris/meulleri/nemorosa/subulata	503-553			
Araucaria laubenfelsii	444-494			
Araucaria luxurians/schmidii	430-480			
Araucaria montana/scropulorum	443-493			
Araucaria rulei	529-579			
Araucaria bernieri/biramulata/columnaris/laubenfelsii/luxurians/meulleri/Montana/nemorosa/rulei/scropulorum/schmidii/subulata	594-602	1.40	AAGGAAATA	SAK

Table S2. cont.

1	2	3	4	5
Araucaria heterophylla	504-554	0.99	CCCCGATATATATATATAGATAGAGATATATATATCTATTACCTGTAGAT	BDGP
Araucaria cunninghamii	219-269	0.97	CAGACTGGGCTATAAACCCAGACGGTAAATCCGTCGTCCCTTTGAGACTA	BDGP
	554-564	1.35	CAAGGAAATAT	SAK
Araucaria humboldtensis	82-132	0.97	CATACTGGGCTATAAACCCAGACGGTAAATCCGTCGTCCCTTTGAGACTA	BDGP
Araucaria hunsteinii	437-487	0.98	CTATTACCTATATATATAGACACGTATCTATACTTTCAAGGAAATATAAG	BDGP
	473-483	1.51	CAAGGAAATAT	SAK
Cycadaceae				
Cycas multipinnata	165-215	0.91	TGGTCATATTAATATATGGGTCTCATATGGCATGGATGCTAGAGATCATC	BDGP
Cycas ophiolitica/platyphylla	138-188	0.89	TGGTCATATTAATATATGGGTCTCATATGGATGGGCATGGATGCTAGAGA	BDGP

Table S3. Tandem repeats found in the *psbA-trnH* sequences of gymnosperms

Taxon	Length(bp)	Sequence	Copy number	Match point	Score	Location
Araucaria bernieri	13	TAAATCTAGACTC	3.9	0.97	93	135-185
Araucaria biramulata/ subulata	13	TAAATCTAGACTC	2.9	0.96	67	135-172
Cycas multipinnata	27	TAAAAAGAAAGGTTTGGTACTCTTCTT	2.0	0.96	101	67-121
Amentotaxus formosana	13	GATTCTATACTAA	1.9	1	50	198-222
Taxus cuspidata/var. nana	95	TATACTATTTAGATATAATATATCTATATATATTATCTTATATTAGGTACCCAATCTGATTCTCTTATTATTCGATTCATGCCTATTGCTTTCAA	3.0	0.98	551	226-514
Taxus canadensis	93	TATACTATTTAGATATAATATATATATATATTATCTTATATTAGGTACCCAATCTGATTCTCTTATTATTCGATTCATGCCTATTGCTTTAAA	3.0	0.98	539	231-513
*Taxus media/ hunnewellian*a	95	TATACTATTTAGATATAATATATCTATATATATTATCTTATATTAGGTACCCAATCTGATTCTCTTATTATTCGATTCATGCCTATTGCTTTCAA	4.0	0.98	741	226-609
Taxus yunnanensis	69	TGACAATTAGACTATAAATAGATATAATATATCTATAGATACCAAAAGAGAGGTTTTTATAATTTGACT	3.5	0.98	470	395-638
Taxus wallichiana	69	TGACAATTAGACTATAAATAGATATAATATATCTATAGATACCAAAAGAGAGGTTTTTATAATTTGACT	2.5	0.98	332	394-568
Microbiota decussata	17	TGAATAATCTAATAGTT	2.1	1	70	353-387

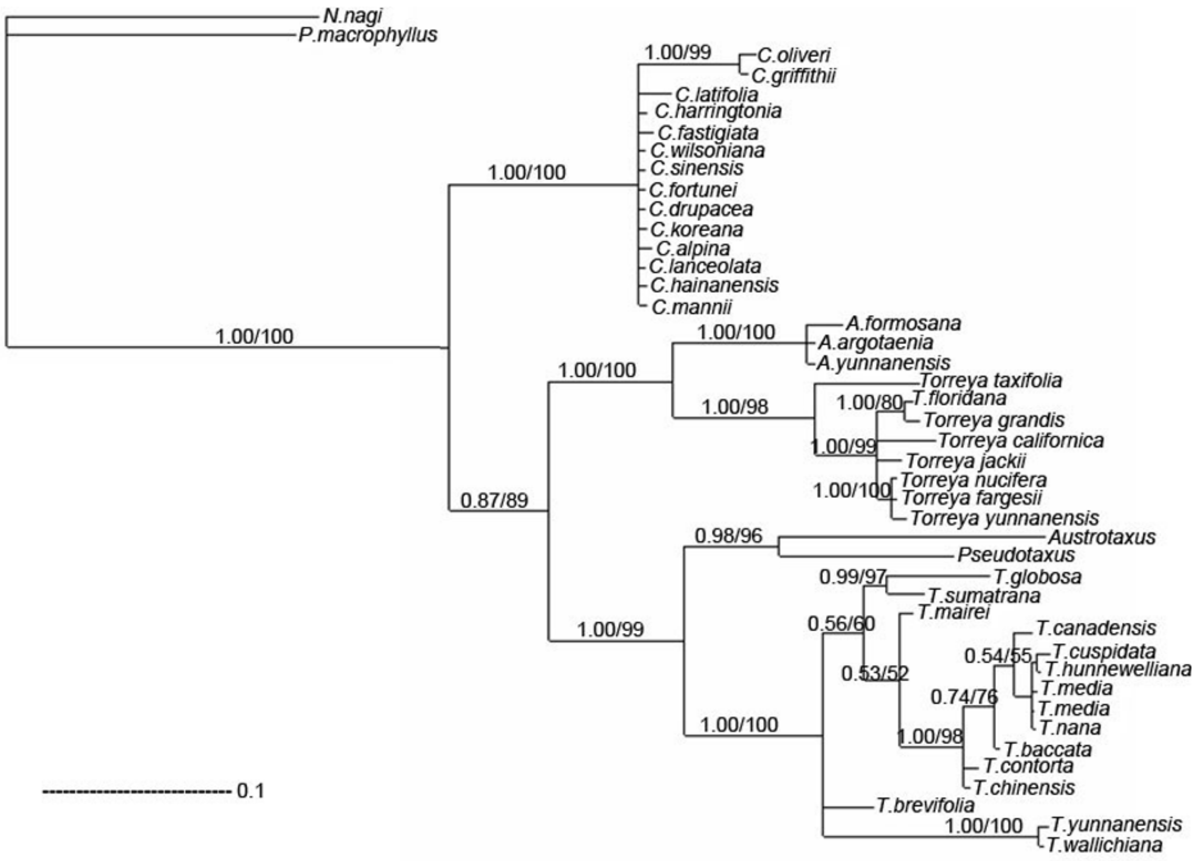


Figure S1. Bayesian 50% majority rule consensus tree (7000 trees sampled; burn-in = 1750 trees) inferred from the Taxaceae and Cephalotaxaceae *psbA-trnH* sequence alignment under the GTR+G model (gamma shape parameter: 0.8005). Bayesian PPs are given beside branches, before slash (/). ML BPs are given after slash. Branch lengths (scale bar, expected number of substitutions per site) are proportional to the mean of PPs of branch lengths of sampled trees.