Phylogenetic Relationships of the Genus Taxus Inferred from Chloroplast Intergenic Spacer and Nuclear Coding DNA

Da Cheng HAO, a,c BeiLi HUANG, b and Ling YANG*, a

a Laboratory of Pharmaceutical Resource Discovery, Biotechnology Division, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian 116023, China; b LuShan Botanical Garden, Chinese Academy of Sciences, JiangXi 332900, China; and c The Graduate University, Chinese Academy of Sciences, Beijing 100049, China.

Received September 10, 2007; accepted November 30, 2007; published online November 30, 2007

A cladistic analysis of the medicinal plant Taxus, using the sequences of one chloroplast (trnS-trnQ spacer) and three nuclear taxadiene synthase (TS), 10-deacetylbaccatin III-10β-O-acetyltransferase (DBAT), and 18S rDNA) molecular markers, was carried out by distance, parsimony, likelihood, and Bayesian methods. Three of the four New World species (T. brevifolia, T. floridana and T. globosa) form a well-supported clade, whereas T. canadensis initially branches—appearing distantly related to both Old World taxa and New World species. In Asia, Taxus chinensis, T. mairei, T. sumatrana and T. wallichiana cluster together and are sister to a clade containing T. baccata and T. contorta. Taxus yunnanensis is more closely related to T. wallichiana than to four other Taxus species in our study from China; T. contorta is closer to the Euro-Mediterranean T. baccata than to the Asian species. This study provides a genetic method for authentication of economically important Taxus species and proposes a robust phylogenetic hypothesis for the genus. Using trnS-trnQ spacer sequences, we were able to distinguish T. mairei from all other species of Taxus. 

Key words Taxus: phylogeny; trnS-trnQ spacer; 10-deacetylbaccatin III-10β-O-acetyltransferase; taxadiene synthase; 18S rRNA

Plants of the genus Taxus are sources of a number of physiologically and pharmacologically active compounds of different classes, especially the anti-cancer paclitaxel and many other taxane derivatives. There are at least ten species in Taxus. The species of Taxus are more geographically than morphologically separable (Fig. 1). There is a large variation in taxane content between the different species and cultivars.1) It is essential to find suitable plants for the various production protocols of paclitaxel. The correct identification of the Taxus species is not only a prerequisite for the relevant plant breeding and selection, and Good Agricultural Practice (GAP), but also a precondition for the chemical and pharmacological investigations of the respective Taxus species, and Good Manufacturing Practice (GMP).2,3) Taxus mairei, an endemic Taxus species of southern China, is the most important source plant for 7-xylosyltaxanes that can be converted to pa-

Fig. 1. Leaf Morphology of some Medicinal Taxus Species
(A) T. baccata; (B) T. canadensis; (C) T. chinensis; (D) T. cuspidata; (E) T. mairei; (F) T. media.

* To whom correspondence should be addressed. e-mail: yling@dicp.ac.cn © 2008 Pharmaceutical Society of Japan
clitaxel and other useful taxanes via chemical transformation and biotransformation. However, it is difficult to authenticate and differentiate *T. mairei* from other source plants of taxanes by morphology. Therefore, molecular studies will definitely be required to resolve the problem.

There is currently debate about the number of species, the classification, and the phylogenetic relationship. Ten, nine, and ten species, and different varieties have been proposed by different authors. Farjon treated *T. mairei* as a variety of *T. chinensis*, while Fu *et al.* treated *T. mairei* and *T. chinensis* as varieties of *T. wallichiana*; the relationship between *T. contorta* and others has not been resolved; Spjut recognized *T. yunnanensis* as a variety of *T. wallichiana*, while it is not in the list of Farjon and Fu *et al.* regarded it as a synonym of *T. wallichiana*. *T. sumatrana* from Indonesia and the Philippines, has been recorded by de Laubenfels; however, *T. sumatrana* is a broad concept which includes the taxa of *T. contorta* and *T. chinensis* and *T. wallichiana var. mairei*. The total number of *Taxus* occurring in Asia is quite controversial and in need of further study.

The objective of our study is to establish genetic profiles for economically valuable medicinal *Taxus* and to provide an assessment of phylogenetic relationships among species in *Taxus* using sequence data from the nuclear (nr) protein-coding genes taxadiene synthase (TS) and 10-deacetyl/baccatin III-10β-O-acetyltransferase (DBAT), 18S rRNA gene, and the chloroplast (cp) non-coding trnS–trnQ intergenic spacer. 18S rRNA sequences have been extensively used for plant molecular phylogenetic analyses, while the secondary metabolism genes TS and DBAT and the cp trnS–trnQ spacer are for the first time used in evaluating relationships within *Taxus* for comparison to morphology-based classifications and to geographic distribution. In paclitaxel biosynthetic pathway, *TS* catalyzes the cyclization of geranylgeranyl diphosphate to geographic distribution. In paclitaxel biosynthetic pathway, for comparison to morphology-based classifications and to the first time used in evaluating relationships within *Taxus* is 18S rDNA, and the phylogenetic relationship. Ten, four, nine, five, and six species, including protein-coding gene, RNA-coding gene, and cp non-coding sequence. In the present molecular phylogeny, we extend the set of available *Taxus* molecular sequences to 21 taxa, representing all species. The present study reports the results of the combined analysis of four molecular markers to differentiate all *Taxus* species.

**MATERIALS AND METHODS**

**Taxon Sampling** Species, geographic origin of the sequenced material, their voucher numbers, and GenBank accession numbers of the sequences generated in this study, as well as those retrieved from GenBank, are given in Table 1. Twenty-five chloroplast and 41 nuclear sequences were newly generated for this study.

**DNA-Extraction, -Amplification, and -Sequencing** Fresh or silica-dried leaves were ground into powder. Genomic DNA was extracted by using Universal Genomic DNA Extraction kit (Takara, Dalian, China), following the manufacturer’s protocol. A 0.9% agarose gel was run to assess the presence and integrity of the DNA. Quantification was done spectrophotometrically and the concentration of DNA ranged from 50—77 ng per µl.

A 50 µl PCR reaction mix consisted of 5 µl of 10× reaction buffer, 4 µl each of 2.5 mM dNTPs stock, 2.5 µl of 10 µM forward and reverse primers (synthesized by Takara, Dalian, China), 0.5 µl bovine serum albumin (10 mg/ml), and 1.5 units of Ex Taq polymerase (Takara, Dalian, China). The primers11 used for amplification of trnS–trnQ are, trnS-GGG, taccaggtttgacccte (forward, 5’−3’) and 5’trnL-GAG-R: tccaggatgccgcatc (reverse). The primers used for amplification of 18S rDNA are, F: tcaagattaagctgcatg and R: caacatcggcgagcaacacc. The TS gene was amplified using atgcctagcctagattc (forward) and gcagccgcaattgaattc.

**Table 1. Voucher Information and GenBank Accession Numbers for Taxa Used in This Study**

<table>
<thead>
<tr>
<th>Species, origin, voucher specimen, and GenBank accession number,</th>
<th>18S rDNA, TS, DBAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species, geographic origin of the sequences generated in this study</td>
<td><em>TS</em>, China, SN001, DQ888590, AY544988*, AY007027*, EU017135,</td>
</tr>
<tr>
<td>25 different species, including protein-coding gene, RNA-coding gene, and cp non-coding sequence. In the present molecular phylogeny, we extend the set of available <em>Taxus</em> molecular sequences to 21 taxa, representing all species. The present study reports the results of the combined analysis of four molecular markers to differentiate all <em>Taxus</em> species.</td>
<td>21 different species, including protein-coding gene, RNA-coding gene, and cp non-coding sequence. In the present molecular phylogeny, we extend the set of available <em>Taxus</em> molecular sequences to 21 taxa, representing all species. The present study reports the results of the combined analysis of four molecular markers to differentiate all <em>Taxus</em> species.</td>
</tr>
</tbody>
</table>

All unmarked vouchers are deposited in Herbarium, LuShan Botanical Garden, Chinese Academy of Sciences, JiangXi, China (LUS). An asterisk indicates a sequence obtained from GenBank, and a dash indicates missing data.
The DBAT gene was amplified using atggcaggtcacaagaggatttgg (forward) and tcaaggtttagttacatatttgtttg (reverse). Approximately 50 ng of genomic DNA was used as a template for the reaction. The reaction mixture was placed in a Takara PCR Thermal Cycler Dice (Takara, Japan). Cycling (38 cycles) condition was described previously.12) The annealing temperatures were 53 °C (for (38 cycles) condition was described previously.12) The annealing temperatures were 53 °C (for 18S rDNA, TS, and trnS–trnQ) or 49 °C (for DBAT). The PCR products were purified by 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. ABI Prism, BigDye Terminator, and cycle Sequencing Ready Reaction Kit (Applied Biosystems). Two PCR products from one sample were sequenced twice in both directions. ABI Prism, BigDye Terminator, and cycle Sequencing Ready Reaction Kit (Applied Biosystems). Two PCR products from one sample were sequenced twice in both directions.

Phylogenetic Analyses Sequence alignment was performed with CLUSTAL W. The aligned chloroplast and nuclear matrix comprised 2538 and 4052 positions, respectively. Each separate DNA region, as well as all combined data, was analyzed with Modeltest 3.813) to find the best model of evolution for the data (Table 2). Employing the Akaike information criterion (AIC), the model with the lowest AIC score was chosen.

Maximum likelihood (ML), Maximum parsimony (MP), and Neighbour joining (NJ) analyses were performed on the separate molecular partitions and on the combined data. ML analysis and bootstrapping were performed using GARLI 0.951.14) GARLI searches relied on the GTR+G, GTR+I+G, and GTR+G models, which ModelTest selected as the best fitting models for unpartitioned cpDNA, nrDNA, and combined data, respectively. MP analysis was performed using PAUP* 4.0b10.15) Heuristic searches were performed using tree bisection–reconnection branch-swapping and 10 random sequence addition replicates. All sites were equally weighted and gaps were treated as missing characters. NJ analysis was performed with MEGA4.16) Using Maximum Composite Likelihood method for estimating evolutionary distances between all pair of sequences simultaneously. Strong support for individual node is defined as nodes with Bayesian posterior probabilities (PP)≥0.95 or non-parametric bootstrap (BP)≥60—70. No strongly supported conflicting relationships were recovered from cp and nr datasets, so all data were combined for further analyses. We did not employ the incongruence length difference test as it has been shown to be a poor test of the compatibility of separate data partitions.17)

The data sets were analyzed in combined mixed-model analyses using MrBayes 3.1.2.18) The analyses of nr, cp, and combined data utilized seven, one, and eight molecular partitions, respectively (Table 2). Two independent runs with one cold and three heated Markov chains per analysis were

Table 2. Number of Aligned and Informative Positions, Base Frequencies, Best-Fit Models, −ln Likelihood with and without Enforcing a Molecular Clock, and Test of Heterogeneity among Lineages for All Loci Analyzed

<table>
<thead>
<tr>
<th>Partition (no. Taxa)</th>
<th>Alignment positions</th>
<th>Base frequency</th>
<th>P&lt;sub&gt;a&lt;/sub&gt;</th>
<th>Best-fit model for Bayesian analysis (AIC)</th>
<th>−ln likelihood</th>
<th>−2 log LR</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>TrnS&lt;sup&gt;GCU&lt;/sup&gt;–TrnQ&lt;sup&gt;UAG&lt;/sup&gt;</td>
<td>Total Inform. (%)</td>
<td>0.359, 0.158</td>
<td>0.9982</td>
<td>GTR+G</td>
<td>10404.84</td>
<td>1959.94</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>18S rDNA</td>
<td>1702</td>
<td>0.242, 0.227</td>
<td>1.0</td>
<td>Trn+I (HKY+I)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>3089.77</td>
<td>156.36</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>DBAT 1st codon</td>
<td>440</td>
<td>0.272, 0.188</td>
<td>1.0</td>
<td>HKY+I</td>
<td>769.68</td>
<td>21.20</td>
<td>&gt;0.05</td>
</tr>
<tr>
<td>DBAT 2nd codon</td>
<td>440</td>
<td>0.284, 0.211</td>
<td>1.0</td>
<td>HKY</td>
<td>770.78</td>
<td>19.04</td>
<td>&gt;0.10</td>
</tr>
<tr>
<td>DBAT 3rd codon</td>
<td>440</td>
<td>0.293, 0.172</td>
<td>1.0</td>
<td>K81uf (GTR)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>998.78</td>
<td>34.52</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>TS 1st codon</td>
<td>348</td>
<td>0.307, 0.183</td>
<td>1.0</td>
<td>HKY</td>
<td>752.29</td>
<td>51.06</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>TS 2nd codon</td>
<td>343</td>
<td>0.294, 0.249</td>
<td>1.0</td>
<td>HKY+G</td>
<td>663.26</td>
<td>35.14</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>TS 3rd codon</td>
<td>343</td>
<td>0.212, 0.228</td>
<td>1.0</td>
<td>TVM+G (GTR+G)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>812.54</td>
<td>95.16</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>All combined</td>
<td>6590</td>
<td>0.297, 0.194</td>
<td>0</td>
<td>Partitioned mixed model</td>
<td>18440.33</td>
<td>677.0</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Data (24)</td>
<td>986</td>
<td>0.227, 0.282</td>
<td></td>
<td></td>
<td>18778.83</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> Significant values were based on χ<sup>2</sup> tests for homogeneity across taxa as implemented in PAUP* 4.0b10. <sup>b</sup> Implemented in MrBayes 3.1.2.
performed simultaneously until the average standard deviation of split frequencies between the two runs dropped below 0.01. Analyses were run twice to check for consistency of results. We ran two simultaneous runs for $3 \times 10^6$ generations and sampled trees every 100 generations. Topology and branch-length information were summarized in 50% majority rule consensus trees; samples obtained before stationarity of $-\ln$ likelihoods against generations had been reached were discarded as burn-in.

RESULTS AND DISCUSSION

*Taxus* is notorious for being taxonomically difficult. Except from the split between three North American species and the other species, phylogenetic relationships within the genus have remained enigmatic for the most part, and classification schemes currently in use do not rest upon well-supported hypotheses about the underlying phylogeny. Due to limited taxon sampling, the molecular studies conducted have so far provided only few detailed insights into relationships within *Taxus*. In Li *et al*.’s *Taxus* study, the controversial *T. sumatrana*, *T. yunnanensis* and *T. wallichiana* were not included and only internal transcribed spacer (ITS) sequences were used to infer the phylogenetic relationship. Collins *et al*. noted that *Taxus* species delimitation remained a problem but their study focused on two hybrids and their parental species. In the present study, we have substantially increased taxonomic sampling of nrDNA and cpDNA for *Taxus* and provide a much more comprehensive picture of their phylogeny.

Sixty-six new sequences were generated for 27 taxa. DNA sequence lengths and general characteristics for each gene and spacer are summarized in Table 2. Final alignments comprised 1030 aligned positions from the *TS* gene, 1320 from the *DBAT* gene, 2538 from the *trnS–trnQ* spacer, and 1702 from 18S rDNA. No significant variation of base frequencies was observed among taxa within each partition (Table 2). *TS* and *DBAT* sequences exhibited no stop codons or frameshift mutations. As expected, 18S rDNA, *DBAT* 1st and 2nd codon positions evolved slower than other partitions and have fewer parsimony informative sites.

The cp and nr trees showed similar relationships among taxa except a lower interspecific resolution in the nr tree (data not shown). All data were therefore combined to have a “total” evidence hypothesis. In the combined tree (Fig. 2), monophyly of *Taxus* was highly supported, and the clade formed by *T. canadensis* and *T. hunnewelliana* UBC200707 was well supported and basal to the rest of *Taxus*. Two hybrids and their parental species *T. cuspidata* were between *T. canadensis* and a large group formed by other species. *T. contorta* and *T. baccata* formed a basal subclade in this large group, while the others fell into two subclades. Three North American species formed a well-supported subclade (PP = 1.00; BP = 100) and four Asian species formed the other (PP = 0.91). Branches within *Taxus* were much shorter than outside branches (Fig. 2), which indicates that they might have undergone a relatively recent radiation. Alternatively, evolutionary rates might have slowed down in *Taxus* after the lineages split.

de Laubenfels designated only a single species of Asian yew where Farjon sees five (*T. chinensis*, *cuspidata*, *contorta*, *sumatrana*, *wallichiana*), all with disjunct distribu-

---

**Fig. 2.** Bayesian 50% Majority Rule Consensus Tree (10000 Trees Sampled; Burn-in=2500 Trees) Inferred from nr+cp DNA Alignment under the Partitioned Model

Bayesian posterior probabilities are given above branches, before slash (/). ML bootstrap proportions (%) calculated under the GTR+G model are given above branches, after slash (—, clade not included in the tree). MP and NJ bootstrap proportions are shown below branches (MP/NJ). Branch lengths (shown on the right; scale bar, expected number of substitutions per site) are proportional to the mean of the posterior probabilities of the branch lengths of the sampled trees.
tions. Luu et al.\(^{21}\) assigned \(T. \) sumatrana as \(T. \) chinensis or \(T. \) wallichiana, on the basis of their geographical locations. All these treatments are open to testing against the molecular evidence. Li et al.\(^{11}\) constructed an MP tree for 10 \(T. \) species and found that three North American species form a well-supported clade, which was also present in our cp and combined trees (Fig. 2 and data not shown). However, a few other clades were only weakly supported. Among them, the clade consisting of \(T. \) cuspidata and two hybrids were also recovered in our cp tree (data not shown), with high PP and BP support. For the grouping of \(T. \) mairei and \(T. \) chinensis, there is conflict between Li et al.’s ITS tree and our phylogenetic trees. In our cp tree, \(T. \) chinensis was more closely related to \(T. \) contorta and \(T. \) baccata than to \(T. \) mairei; in the combined tree, \(T. \) mairei was more closely related to \(T. \) sumatrana and \(T. \) wallichiana than to \(T. \) chinensis. Due to high level of homoplasy of ITS marker and the MP method used in Li et al.’s study, their results might be less reliable.

Three \(T. \) species were included in Li et al.’s study\(^{10}\) with \(rbcL \) and \(trnL–trnF \) spacer sequences (\(trnL \) intron and exon not included). \(T. \) mairei was shown to be closer to \(T. \) chinensis than to \(T. \) yunnanensis, which is in accordance with our cp tree (data not shown). \(T. \) chinensis and \(T. \) mairei (Fig. 1) can be easily differentiated from \(T. \) yunnanensis through the texture of their leaves. The leaves of the former two are thick, while the latter, are thin and soft. In any case, our phylogeny does not support to treat \(T. \) chinensis and \(T. \) mairei as varieties of \(T. \) wallichiana.

Spjut\(^{8} \) classified \(T. \) into three groups: the \(Bacatta \) group, represented by \(T. \) canadensis, \(T. \) baccata, \(T. \) caespitosa, \(T. \) biternata, \(T. \) cuspidata, \(T. \) recurvata, and \(T. \) umbraculifera that also include varieties, the \(Wallichiana \) group, represented in North America by \(T. \) globosa var. flordiana, \(T. \) globosa var. globosa, and \(T. \) brevifolia, and in Asia by \(T. \) wallichiana, \(T. \) florinii, and \(T. \) suffnessii, and he further distinguished a \(Chinensis \) subgroup in Asia representing six species, three of which were published (\(T. \) chinensis, \(T. \) obscura, \(T. \) phytonii), occurring from central China to Indonesia, and a \(Sumatrana \) group represented by four species, \(T. \) celebica, \(T. \) mairei, \(T. \) kingstonii, and \(T. \) sumatrana with a number of varieties. These morphological groups receive some support from our molecular phylogeny (Fig. 2). On the other hand, results of the chemical analyses of six neutral taxanes allowed the distinction of \(T. \) brevifolia, a North American (\(T. \) canadensis, \(T. \) flordiana, and \(T. \) globosa) and a Eurasian (\(T. \) baccata, \(T. \) wallichiana, \(T. \) cuspidata, and two hybrids) group of \(T. \) species,\(^{22} \) which does not perfectly match our molecular phylogeny. It is known that most species are highly variable in taxoid content and we discuss species that have a similar chemotype. Based on our results—employing multiple molecular markers and a sampling of species in three species groups—we find that there are interspecies relationships within each group that may have profound nomenclatural and taxonomic implications.

A six base pair deletion (AACTGG) that was observed at alignment positions 1666—1671 in the \(trnS–trnQ \) spacer of \(T. \) mairei (Fig. 3), and alignment positions 370—372 (GAA for \(T. \) mairei), 1446 (C for \(T. \) mairei), 1538 (T for \(T. \) mairei), and 1555 (A for \(T. \) mairei) differentiate \(T. \) mairei unambiguously from other species, which can be used to design diagnostic PCR primers. A four base pair deletion at positions 830—833 of \(T. \) chinensis and a six bp insertion at positions 1120—1125 of \(T. \) walllichiana and \(T. \) yunnanensis could be utilized as molecular markers for the respective species. Moreover, combinations of base pair changes at additional positions shown in Fig. 3 could be utilized for distinguishing among the remaining taxa.

We are dealing with an important group of medicinal plants with very meager knowledge of their authentication and interspecific relationships. Our study is by far the most comprehensive molecular analysis of \(T. \) conducted to date, demonstrating that the existing ‘species’-level classification of yew is probably largely artificial. Our molecular data are of great help to identify the botanical origin of the \(T. \) plant and to evaluate systematic relationships. However, our results await corroboration by analyses of more nuclear and/or mitochondrial protein-coding genes.

Acknowledgements We thank the following experts for providing plant materials: YunFen Geng (YunNan Academy of Forestry, KunMing, China), YinKe Zhang (HangZhou Botanical Garden, China), Ron Determann (Atlantic Botanical Garden, GA, U.S.A.), Christopher Quinn (Royal Botanic Garden, Sydney, Australia), Robert G. Nicolson (Smith College, U.S.A.), Stephane Bailleul (Montreal Botanical Garden, Canada), Eric La Fountaine (University of British Columbia)
Botanical Garden, Canada), and James Stevenson (University of Oxford Botanic Garden, U.K.). This study was supported by the National 973 Project (2007CB707802) of the Department of Science & Technology of China, the Innovation Fund of Drug R & D (2007E11SF0052) of Dalian Bureau of Science & Technology, and the Innovation Fund of Dalian Institute of Chemical Physics, Chinese Academy of Sciences.

REFERENCES