Transcriptome Analysis of *Taxus cuspidata* Needles Based on 454 Pyrosequencing

Authors Qiong Wu[1], Chao Sun[1], Hongmei Luo[1], Ying Li[1], Yunyun Niu[1], Yongzhen Sun[1], Aiping Lu[2], Shilin Chen[1]

Affiliations
[1] Institute of Medicinal Plant Development (IMPLAD), Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, P. R. China
[2] Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences, Beijing, P. R. China

Key words
- transcriptome analysis
- *Taxus cuspidata*
- Taxaceae
- 454 pyrosequencing
- Taxol biosynthesis
- transcription factor

Abstract

Taxus species are highly valued as renewable resources for the production of Taxol. Despite the commercial and medicinal importance of *Taxus*, little genomic information is available for yew species, and Taxol biosynthesis still needs to be fully elucidated. In this study, 454 pyrosequencing technology was employed to produce an expressed sequence tag (EST) from the needles of *Taxus cuspidata*. In all, 81 148 high-quality reads from the needles of *T. cuspidata* were produced using Roche GS FLX Titanium. A total of 20 557 unique sequences were obtained, including 12 975 singletons and 7582 contigs. Approximately 14 095 unique sequences were annotated by a similarity search against five public databases. Gene ontology revealed 11 220 unique sequences that could be assigned to 45 vocabularies. In the Kyoto Encyclopedia of Genes and Genomes mapping, 2403 transcripts were established as associated with 3821 biochemical pathways. Enzymes in the plastidial 2-C-methyl-D-erythritol 4-phosphate pathway were well represented. Candidates of the putative genes of Taxol biosynthesis were revealed, including those in the remaining steps. In total, 291 transcripts were identified, representing putative homologues of transcription factors. Furthermore, 753 simple sequence repeat motifs, which are potential molecular markers for genetic application, were identified. These results provide the largest EST collections in *Taxus* and will contribute to biosynthetic and biochemical studies that lead to drug improvement.

Abbreviations

cDNA:	complementary DNA
EST:	expressed sequence tags
GO:	gene ontology
KEGG:	Kyoto Encyclopedia of Genes and Genomes
SSR:	simple sequence repeat
TFs:	transcriptional factors

Supporting information available online at http://www.thieme-connect.de/ejournals/toc/plantamedica

received April 25, 2010
revised July 22, 2010
accepted August 14, 2010

Bibliography
DOI http://dx.doi.org/10.1055/s-0030-1250331
Published online September 22, 2010
Planta Med 2011; 77: 394–400

ISSN 0032-0943

Correspondence
Prof. Dr. Shilin Chen
Institute of Medicinal Plant Development (IMPLAD)
Chinese Academy of Medical Sciences & Peking Union Medical College
No. 151, Malianwa North Road, HaiDian District
Beijing 100193
People's Republic of China
Phone: + 86 10 62 89 97 00
Fax: + 86 10 62 89 97 76
slchen@implad.ac.cn

Introduction

Taxus cuspidata (Taxaceae) is one of the most important medicinal plants in Northeast Asia. It has received much of the world's attention in recent years because it contains Taxol, a well-known anti-cancer agent, in its needles and bark. *T. cuspidata* is a long-lived, shade-tolerant, dioecious tree in the conifer genus *Taxus*. As a tertiary relic species, it is native to China, Korea, Japan, and Eastern Siberia [1]. In China, it has long been used for treating diabetes, promoting diuresis, and as an emmenagogue.

Taxol biosynthesis in yew species involves the cyclization of geranylgeranyl diphosphate, the universal diterpenoid progenitor. This parental oletin is then functionalized by a series of cytochrome P450-mediated oxygenations, CoA-dependent acyl/aroyl transfers, oxidation at C9, and oxetane formation, yielding the intermediate baccatin III, to which the functionally important C13-side chain is appended [2–8]. In *Taxus* species, the isoprenoid precursor comes from the plastid-localized plastidial 2-C-methyl-D-erythritol 4-phosphate (MEP) pathway [9]. Over 350 taxane diterpenoids have been discovered in the yew species, and Taxol is not the main metabolite. A very large number of side-chain variants has been isolated; these variants differ in the position of the hydroxylated taxane nucleus as well as the type of acyl/

aroyl substitution, and greatly affect the yield of Taxol production [8, 10]. At present, the biosynthesis pathway of Taxol is not fully elucidated and its taxoids transcriptional regulation is not clear. All these hamper improvements in drug production.

Expressed sequence tag (EST) represents a rapid and cost-effective method for analyzing transcribed genome regions. It identifies the genes involved in plant secondary metabolism [11]. EST analysis is also a powerful approach for discovering new genes and analyzing gene expression profiles in tissues [12]. The advent of next-generation sequencing technologies has led to a revolution in genomics and genetics, and provided a cheaper and faster delivery of sequencing information. The 454 pyrosequencing is the most widely used new platform in *de novo* sequencing and is applied in non-model organisms including medicinal plants [13, 14].

Despite the commercial and medicinal importance of *Taxus* species, there is little genomics or transcriptome research on these species. Only 179 EST can be retrieved freely from the GenBank. In this study, through the Roche GS FLX system, 81 148 high-quality reads were obtained from a cDNA library generated from *T. cuspidata* needles. Bioinformatics analysis indicated that all genes in the plastidial 2-C-methyl-D-erythritol 4-phosphate pathway and eight genes involved in Taxol biosynthesis exist in the transcripts of yew needles. Furthermore, a few candidate genes that encode putative enzymes of the remaining steps in Taxol pathway, the putative transcriptional factors, and EST-derived SSR markers were screened. The gene expression profiles in this tissue were studied. The transcriptome analysis of *Taxus* species based on a next-generation sequencer reported in this study is the first of its kind.

Materials and Methods

Plant material

Detailed descriptions of plant material are included as Supporting Information (**Text 1, Fig. 1S,** Supporting Information).

RNA extraction and cDNA library construction

RNA extraction was performed according to the standard protocol of RNeasy Plant Kit, and cDNA was synthesized using the SMART™ PCR cDNA Synthesis Kit (**Text 2**, Supporting Information).

454 Library preparation and sequencing

For 454 sequencing, amplified cDNAs were sheared by nebulization to produce random fragments (**Text 3**, Supporting Information).

Sequence assembly and annotation

For detailed sequence assembly and annotation, refer to Supporting Information (**Text 4**, Supporting Information).

Sequence comparisons with other species

Similarity searches were performed against the Dana-Farber Cancer Institute gene indices available for oryza (OGI17.0), pine (PGI7.0), poplar (PPLGI4.0), and spruce (SGI3.0) (**Text 5**, Supporting Information).

Gene ontology (GO) analysis and metabolic pathway assignment using the Kyoto Encyclopedia of Genes and Genomes (KEGG)

The unique sequences were assigned GO terms based on the top BLAST hit for queries in the *Arabidopsis* proteins and cDNA sequences (TAIR9). Metabolic pathway assignments were carried out according to the KEGG resource (**Text 6**, Supporting Information).

Identification of simple sequence repeat (SSR) and Sanger sequencing validation

Detailed descriptions are discussed in Supporting Information (**Text 7**, Supporting Information).

Supporting information

Detailed descriptions of materials and methods, GO categories, KEGG mappings, candidate genes of Taxol biosynthesis, sequences with putative functions similar to transcription factors, occurrence and number of repeated units of SSR motifs, and primers used in SSR validation through Sanger methods are available as Supporting Information.

Results and Discussion

Recently, high-throughput sequencing technologies have been developed rapidly and applied widely in genomic study of many organisms. For instance, 454 pyrosequencing has been used in *de novo* sequencing and has made EST-derived resources more readily accessible for non-model organisms [11, 13]. In this study, a SMART cDNA library was constructed from the total RNA extracted from the needles of *T. cuspidata*. In all, 81 148 high-quality reads were produced with an average length of 389 bp and an average quality score of 31. Then, 80 028 reads that passed high-quality filtering and a minimum of 50 contiguous nucleotides were used in the assembly (**Table 1**). The *de novo* assembly resulted in 20 557 unique sequences, including 7582 contigs and 12 975 singletons. The sizes of the unique sequences ranged from 50 to 3555 bp with an average length of 456.6 bp (**Fig. 1**). Thus,

Table 1 Summary of 454 pyrosequencing of *T. cuspidata* cDNA library.

Descriptive category	Number of 454 ESTs	Length of bases (bp)	%
Total high quality (HQ) reads	81 148	31 604 969	
Reads used in assembly	80 028	31 570 074	
▸ reads assembled in contigs	74 209		
▸ number of contigs	7582	4 476 418	36.9
▸ number of singletons	12 975	4 909 978	63.1
Deeps on contigs	5.34		
Redundancy			92.7
Number of unique sequences	20 557	9 386 396	
▸ average length of unique sequences		456.6	
EST group matches with *e*-value < 10^{-5}			
▸ SwissProt	7916		38.5
▸ Nr	12 069		58.7
▸ *Arabidopsis thaliana*	12 219		59.4
▸ pine	12 169		59.2
▸ spruce	12 957		63.0
▸ no matches	6462		31.4

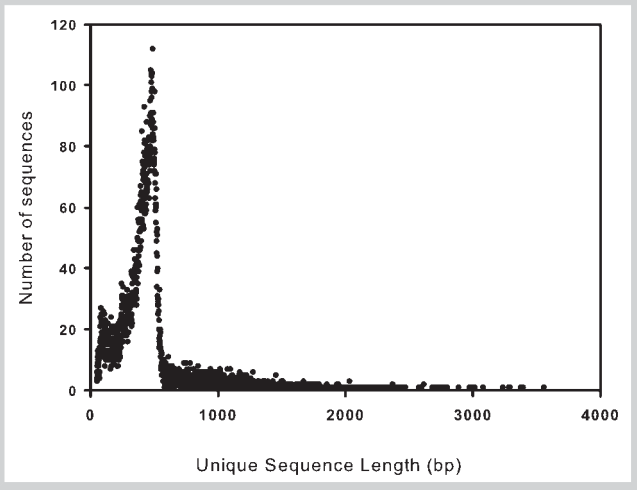

Fig. 1 Size distribution of *T. cuspidata* unique sequences based on 454 pyrosequencing.

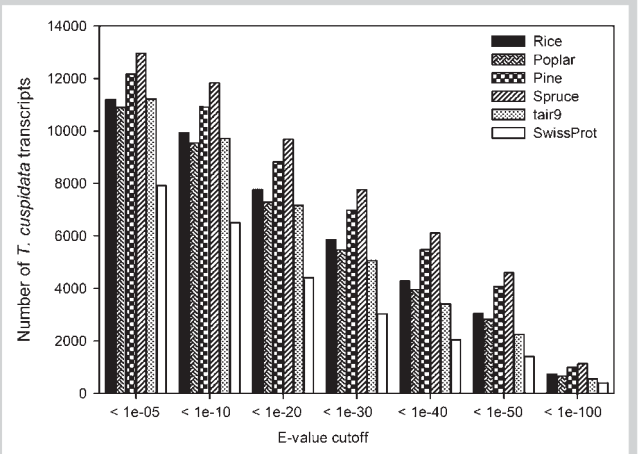

Fig. 2 Transcript sequences of *T. cuspidata* similar to SwissProt, *A. thaliana*, rice, poplar, pine, and spruce databases according to BLASTX or tBLASTX e-value cutoffs.

74209 reads were assembled into contigs with EST numbers ranging from 2 to 2051. An estimated 92.7% of the sequences appeared at least twice among the EST. Comparing the unique sequences against the SwissProt and Nr databases, 38.5 and 58.7% had significant matches in the databases, respectively (E ≤ 1.0e-5). Unique sequences were also compared to gymnosperm EST by tBLASTX at the same e-value, among which spruce and pine showed homologues of 63% (12957) and 59.2% (12169), respectively. A total of 14095 (68.6%) unique sequences shared a significant similarity to the sequences in the above mentioned databases and have orthologs in other organisms. In this study, overall sequence similarity was found to be lower with angiosperms than with gymnosperms and had the largest overlaps with EST in the Spruce Gene Index (SGI3.0) (**Fig. 2**). These results demonstrate that datasets derived from angiosperms are not sufficient for annotation of *T. cuspidata*, and more data should be accumulated for gymnosperms. More than 31.4% (6462) sequences were unidentified in public databases, indicating that novel information was discovered in our 454 datasets. All high-quality reads have been deposited in the GenBank under the accession number SRX015126.

In the cDNA library of *T. cuspidata*, more than 83.5% contigs have 2 to 10 sequences, 9.3% have 11 to 20 sequences, 2.7% have 21 to 30 sequences, and 4.6% have 31 to 2051 sequences (**Table 1S**, Supporting Information). A total of 75 transcripts with reads of at least 102 were abundantly expressed. The most highly expressed genes were coded for ribulose bisphosphate carboxylase (RuBisCo), which assembled three contigs having a total of 3020 EST. In plants, RuBisCo is more abundant during the day when it is transcriptionally regulated by the light receptor phytochrome, thus fixing carbon dioxide in photosynthesis [15]. The genes encoding photosystem I reaction center subunit, chlorophyll a–b binding protein, ribulose-phosphate 3-epimerase, and photosystem II protein, glutamine synthetase, granule-bound starch synthase, caffeoyl-CoA *O*-methyltransferase, sedoheptulose-1,7-bisphosphatase, and metallothionein-like protein were relatively abundant. Glutamine synthetase catalyzes the critical incorporation of inorganic ammonium into the amino acid glutamine and increases photosynthesis and growth at low nitrogen concentrations [16]. Granule-bound starch synthase is responsible for amylose synthesis [17]. Caffeoyl-CoA *O*-methyltransferase participates in lignin biosynthesis in the formation of cell wall ferulic esters [18]. Edoheptulose-1,7-bisphosphatase functions at a branch point in the Calvin cycle between regeneration of the acceptor molecular ribulose 1,5-bisphosphate and exports to starch biosynthesis [19]. Owing to its capability to sequester excess amounts of metal ions, metallothionein participates in homeostasis and antioxidant functions [20]. The transcripts associated with fundamental life processes, such as photosynthesis or energy, significantly affect redundancy in all the EST collections.

In total, 11220 (54.6%) unique sequences were assigned to one or more of the 45 GO categories for cellular component, biological process, and molecular function (**Fig. 2S**, Supporting Information). In the cellular component, chloroplast (23%), plastid (10.7%), plasma membrane (9.3%), and nucleus (8.4%) were well represented. In the biological process, other cellular processes (39%), other metabolic processes (35.6%), protein metabolism (11.5%), abiotic or biotic stimulus (10.7%), and response to stress (10.2%) subclasses were abundant. Among molecular functions, a huge incidence of other enzyme activities (16.7%), other binding activities (12.5%), transferase (12.2%), hydrolase (10.7%), nucleotide binding (8.2%), and protein binding (8.0%) were annotated. Transferase participated in many taxoid metabolism processes and was closely related to taxoid biosynthesis and side routes. In KEGG, a total of 11291 (54.9%) unique sequences found similarities (**Table 2S**, Supporting Information). Among them, 8668 were sorted as unassigned, which had significant similarities to sequences in KEGG but unclear functions. Only 2623 (12.8%) were assigned to a clear functional group: 2403 were mapped to a total of 3821 KEGG biochemical pathways while 220 were only general functions predicted. More than half (60.4%) of the sequences that mapped to specific KEGG pathways were included in basic metabolism processes involved in carbohydrate, amino acid, energy, and lipid metabolism. In secondary metabolism subclass, most transcripts were mapped to phenylpropanoid biosynthesis, carotenoid biosynthesis, limonene and pinene degradation, streptomycin biosynthesis, flavonoid biosynthesis, novobiocin biosynthesis, and terpenoid biosynthesis. Functional annotation by GO and KEGG suggests that the results of the current study

may allow for the identification of novel genes involved in the pathways of secondary metabolite synthesis.

All genes encoding the seven enzymes in the plastidial MEP pathway were identified in the 454 datasets (**Table 2**). The sequences encoding 4-hydroxy-3-methylbut-2-enyl diphosphate reductase (HDR), with 69 EST, were substantially abundant. The two enzymes catalyzing the slow steps of the mevalonate-independent pathway, 1-deoxy-D-xylulose-5-phosphate synthase (DXPS) and 1-deoxy-D-xylulose-5-phosphate reductoisomerase (DXR), held 23 EST each [21,22]. IPP isomerase (IDI) and geranylgeranyl pyrophosphate synthase (GGPPS) held six and five acquisitions, respectively. The reaction catalyzed by IDI transforms the relatively inactive IPP into the active isomer DMAPP and plays a pivotal role in isoprenoids metabolism. DXR and GGPPS were quite upregulated in methyl jasmonate-induced cells [4]. Among the genes involved in Taxol biosynthesis, eight were identified in the 454 datasets. As an enzyme in early steps, taxane 13α-hydroxylase with 34 EST was well represented. Taxane 13α-hydroxylase catalyzes the production of taxadien-5α,13α-diol. Taxane 2α-*O*-benzoyltransferase (TBT), involved in the production of 10-deacetylbaccatin III in the middle step, held 33 EST. Although the biosynthesis of Taxol and related taxoids are beginning to be understood, many steps still remain to be addressed. The C1β-hydroxylation and the C4,C20β-epoxidation leading to oxetane ring formation, oxidation of the C9-hydroxyl group to the corresponding carbonyl in the taxane core formation, CoA-dependent ligation to β-phenylalanine, 2′-hydroxylation of *N*-debenzoyl-2′-deoxytaxol, and *N*-benzoylation of *N*-debenzoyltaxol in the side-chain assembly are the remaining steps [4,23,24]. Available evidence suggests that C4,C20β-epoxidation occurs at the level of a taxadien-hexaol and is followed by ring expansion toward oxetane construction [23]. These steps could be mediated by CYP450. In the side-chain assembly, the catalyzed transformation of 2′-hydroxylation of β-phenylalanoyl baccatin III (*N*-debenzoyl-2′-deoxytaxol) into phenylisoserinoyl baccatin III (*N*-debenzoyltaxol) implies the involvement of CYP450 [25]. Meanwhile, Taxol biosynthesis competes with side routes produced by acyl/aroyl or the oxidation of the taxane nucleus derived from common precursors. Identification of these side-route genes could have an important implication in eventually increasing Taxol yields [26]. For this purpose, we screened our 454 datasets and discovered candidate genes of CYP450, epoxidase, CoA ligases, and *N*-ben-

Table 2 Summary of candidate genes involved in Taxol biosynthesis.

Candidates	Organism	E-value	Unigene	Reads	Accession No.
1-Deoxy-D-xylulose-5-phosphate synthase, DXS	*Capsicum annuum*	0	contig02426	15	O78328
1-Deoxy-D-xylulose-5-phosphate synthase, DXS	*Oryza sativa subsp. japonica*	4.00E-41	contig04660	2	O22567
1-Deoxy-D-xylulose-5-phosphate synthase, DXS	*Oryza sativa subsp. japonica*	1.00E-75	contig06207	6	O22567
1-Deoxy-D-xylulose 5-phosphate reductoisomerase, DXR	*Oryza sativa subsp. japonica*	0	contig04157	23	Q8W250
2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase, MCT	*Arabidopsis thaliana*	9.00E-88	contig06159	9	P69834
4-Diphosphocytidyl-2-C-methyl-D-erythritol kinase, CMK	*Solanum lycopersicum*	1.00E-125	contig05758	14	P93841
2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase, MCS	*Catharanthus roseus*	5.00E-79	contig06336	10	Q9M4W3
2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase, MCS	*Arabidopsis thaliana*	5.00E-34	contig06498	2	Q9CAK8
4-Hydroxy-3-methylbut-2-en-1-yl diphosphate synthase, HDS	*Protochlamydia amoebophila*	1.00E-122	contig04034	25	Q6MD85
4-Hydroxy-3-methylbut-2-enyl diphosphate reductase, HDR	*Cyanothece sp.*	1.00E-147	contig00845	69	B1WTZ2
Isopentenyl-diphosphate Delta-isomerase II, IDI	*Clarkia breweri*	1.00E-104	contig01569	6	Q39471
Geranylgeranyl pyrophosphate synthetase, GGPPS	*Capsicum annuum*	5.00E-25	FXAT9O003CUPVX	1	P80042
Geranylgeranyl pyrophosphate synthetase, GGPPS	*Catharanthus roseus*	8.00E-48	FXAT9O003CZ7X0	1	Q42698
Geranylgeranyl pyrophosphate synthetase, GGPPS	*Cyanophora paradoxa*	1.00E-23	contig06297	3	P48368
Taxadiene 5-α hydroxylase, T5OH	*Taxus chinensis*	9.00E-32	FXAT9O003C72PY	1	gb|AAU93341.1
Taxadiene 5-α hydroxylase, T5OH	*Taxus chinensis*	6.00E-36	FXAT9O003DJ76T	1	gb|AAU93341.1
Taxadiene 5-α hydroxylase, T5OH	*Taxus chinensis*	5.00E-23	FXAT9O003DP38Y	1	gb|AAU93341.1
Taxadiene 5-α hydroxylase, T5OH	*Taxus cuspidata*	3.00E-54	contig00287	13	gb|AAQ56240.2
Taxadien-5-α-oL *O*-acetyltransferase, TAT	*Taxus chinensis*	5.00E-42	FXAT9O003DJBQB	1	Q8S9G6
Taxadien-5-α-oL *O*-acetyltransferase, TAT	*Taxus cuspidata*	8.00E-18	FXAT9O003DO37Z	1	Q9M6F0
Taxane 10-β-hydroxylase, T10OH	*Taxus cuspidata*	3.00E-30	FXAT9O003DLKX1	1	Q9AXM6
Taxane 10-β-hydroxylase, T10OH	*Taxus cuspidata*	3.00E-45	FXAT9O003DQNGU	1	Q9AXM6
Taxane 10-β-hydroxylase, T10OH	*Taxus cuspidata*	3.00E-23	contig05275	4	Q9AXM6
Taxane 10-β-hydroxylase, T10OH	*Taxus cuspidata*	2.00E-99	contig05280	4	Q9AXM6
Taxane 13-α-hydroxylase, T13OH	*Taxus cuspidata*	1.00E-46	FXAT9O003C390L	1	Q8W4T9
Taxane 13-α-hydroxylase, T13OH	*Taxus cuspidata*	4.00E-36	FXAT9O003DJ76T	1	Q8W4T9
Taxane 13-α-hydroxylase, T13OH	*Taxus cuspidata*	2.00E-17	FXAT9O003DP38Y	1	Q8W4T9
Taxane 13-α-hydroxylase, T13OH	*Taxus cuspidata*	8.00E-68	contig00283	12	Q8W4T9
Taxane 13-α-hydroxylase, T13OH	*Taxus cuspidata*	6.00E-52	contig00287	13	Q8W4T9
Taxane 13-α-hydroxylase, T13OH	*Taxus cuspidata*	3.00E-16	contig03175	5	Q8W4T9
Taxane 13-α-hydroxylase, T13OH	*Taxus cuspidata*	4.00E-36	FXAT9O003DJ76T	1	Q8W4T9
Taxoid 2-α-hydroxylase, T2OH	*Taxus chinensis*	2.00E-33	FXAT9O003DDV1Y	1	gb|AAV54171.1
Taxoid 2-α-hydroxylase, T2OH	*Taxus canadensis*	3.00E-39	contig01631	4	gb|AAS89065.2
2-α-Hydroxytaxane 2-*O*-benzoyltransferase, TBT	*Taxus cuspidata*	1.00E-36	FXAT9O003DFUP4	1	Q9FPW3
2-α-Hydroxytaxane 2-*O*-benzoyltransferase, TBT	*Taxus cuspidata*	5.00E-12	FXAT9O003DPL1Z	1	Q9FPW3
2-α-Hydroxytaxane 2-*O*-benzoyltransferase, TBT	*Taxus cuspidata*	2.00E-61	contig00059	31	Q9FPW3
10-Deacetylbaccatin III 10-*O*-acetyltransferase, DABT	*Taxus cuspidata*	3.00E-08	FXAT9O003DBW9Z	1	Q9M6E2
10-Deacetylbaccatin III 10-*O*-acetyltransferase, DABT	*Taxus cuspidata*	5.00E-19	FXAT9O003DJZ16	1	Q9M6E2
3′-*N*-debenzoyl-2′-deoxytaxol *N*-benzoyltransferase, DBTNBT	*Taxus canadensis*	9.00E-28	FXAT9O003DE74B	1	Q8LL69
3′-*N*-debenzoyl-2′-deoxytaxol *N*-benzoyltransferase, DBTNBT	*Taxus canadensis*	3.00E-20	FXAT9O003DOP51	1	Q8LL69

zoyltransferase (**Table 3S**, Supporting Information). These sequences provide useful information for elucidating the remaining steps and exploring the side-route enzymes in Taxol biosynthesis pathway. The proposed biosynthesis pathway of Taxol is summarized in **Fig. 3**.

Transcription factors (TFs) are sequence-specific DNA-binding proteins that interact with the promoter regions of target genes and modulate the rate of initiation of mRNA synthesis by RNA polymerase II [27]. Based on similarity searches, 291 unique sequences were identified representing putative transcription factors, covering zinc finger-containing proteins, WD or WD-40, homeobox, WRKY, MYB (Myble), basic helix-loop-helix (bHLH), auxin response factors (ARFs), NAC domain-containing protein, and ethylene-responsive transcription factor (**Table 4S**, Supporting Information). The zinc finger-containing proteins were the most abundant class in our results with 80 transcripts, including CCCH, CONSTANS-like proteins, DNA binding with one finger (Dof), A20/AN1, zinc finger protein MAGPIE, RING and CHY zinc finger domain-containing protein, among others. CCCH-type proteins containing tandem zinc-binding motifs are characterized by three cysteines followed by one histidine ($CX_{7-8}CX_5CX_3H$) [28]. Such proteins are widely present in eukaryotes, from yeast to mammals, but little is known about their functions in higher plants [29]. CONSTANS-like proteins are considered photoperiod regulators and play a pivotal role in circadian regulation, initiating flowering in response to changes in daylength [30]. Dof domain proteins are plant-specific transcription factors with a highly conserved DNA-binding domain and play diverse roles in gene regulation processes restricted to plants [31]. The A20/AN1 proteins could represent common elements of stress response in plants, and the characterization of such genes presents significant implications on stress tolerance improvement in adversity [32]. Putative homeobox and WRKY were detected in 28 and 18 unique sequences, respectively. WRKY has many regulatory roles in response to biotic and abiotic stresses, and some WRKY proteins have been found to be involved in the signal transduction pathway and defense response against pathogenic attack [33]. Other classes more abundant in our datasets include MYB, bHLH, ARF and NAC domain-containing protein, ethylene-responsive, nuclear transcription factor, heat stress transcription factor, MADS box, and GATA TFs. The molecular functions of these transcription factors vary in plants, with most of them still unclear.

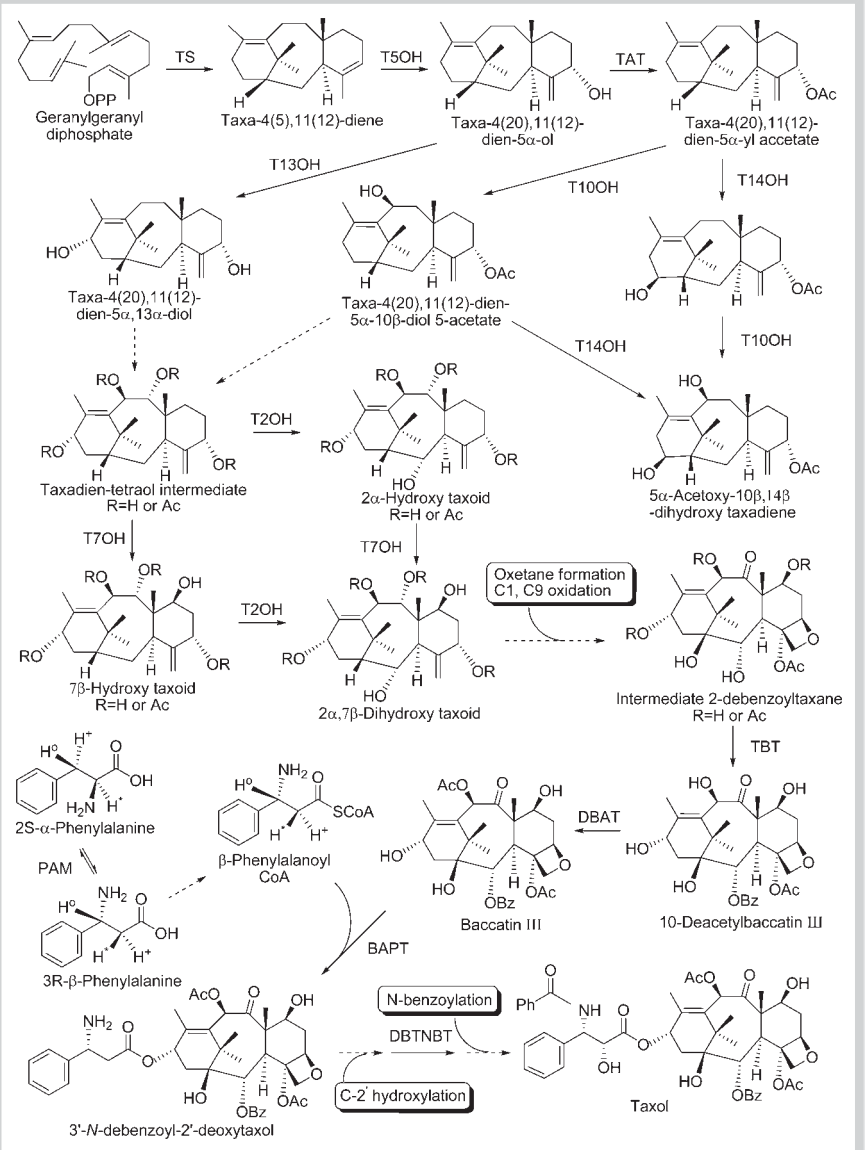

Fig. 3 Putative pathway for the biosynthesis of Taxol in *Taxus*. Abbreviations: BAPT: baccatin III: 3-animo-3-phenylpropanoyltransferase; DBAT: 10-deacetylbaccatin III-10-*O*-acetyltransferase; DBTNBT: 3′-*N*-debenzoyl-2′-deoxytaxol *N*-benzoyltransferase; DMAPP: dimethylallyl diphosphate; GGPPS: geranylgeranyl diphosphate synthase; IPP: isopentenyl diphosphate; PAM: phenylalanine aminomutase; TAT: taxadien-5α-ol-*O*-acetyl transferase; TBT: taxane 2α-*O*-benzoyltransferase; TS: taxa-4(5),11(12)-diene synthase; T2OH: taxoid 2α-hydroxylase; T5OH: cytochrome P450 taxadiene 5α-hydroxylase; T7OH: taxoid 7β-hydroxylase; T10OH: taxane 10β-hydroxylase; T13OH: taxane 13α-hydroxylase.

These results, however, provide information for those probably involved in environmental responses, stress tolerance, or biosynthesis of secondary metabolites in *Taxus*.

753 SSR motifs were identified from 641 SSR-containing unique sequences in the datasets (**Table 5S**, Supporting Information). Approximately 88% of the SSR were between 14 and 24 bp in length. The density was approximately one SSR in 12.47 kb of coding sequences, and the frequency was 3.1%. A total of 317 motif types were identified. The number of repeats for motif units was between 3 and 50, with dinucleotide TA/TA having the highest number of repeat units and the longest motif length of 100 bp. Among SSR-containing unique sequences, the majority (556, 86.7%) had a single SSR per sequence, while 85 (13.3%) contained two or more SSR per sequence. The top 10 motifs were composed of dimers, trimers, tetramers, and pentamers. At 260, hexamers accounted for 34.5% in the population, followed by trimers which accounted for 32.1% (242). Both hexamers and trimers comprised a little more than half of the total microsatellite sequences. The number of repeat units for each motif varied between 7 and 39 times in most cases, but had the highest repeat number of 50 in the unique sequence FXAT9O003DDGKH. Similarity search showed that this sequence had no hits in public databases, which may represent a novel gene in *Taxus*. The most represented repeat types in dicotyledonous species, AG, AT, AAG, AGG, and AGC, were well represented in this study with a total of 72 SSRs. Conversely, CCG, the most common repeated motif in monocotyledonous species, was quite rare. This supports the idea that the SSR composition in gymnosperms is more similar to that of dicots than monocots [34]. The presence of a majority of hexamers and trimers in *T. cuspidata* unique sequences was in agreement with the observation that trinucleotides were significantly more frequent in the coding sequences of *A. thaliana*. Indeed, a mutation in a mononucleotide, dinucleotide, tetranucleotide, or pentanucleotide in the coding region may result in a frameshift that could change translated protein structure and function [34]. To evaluate error rates of SSR detection based on pyrosequencing, 10 SSRs were randomly selected and subjected to Sanger sequencing (**Table 3**). Results showed that Sanger sequencing was successful for eight SSRs, seven of them being the same as those identified by pyrosequencing. The results demonstrate that 454 sequencing can be an efficient genotyping platform, providing an alternative to classic genotyping methods given appropriate application.

Table 3 Simple sequence repeats (SSR) validation by Sanger sequencing.

No.	Unique sequences	454 sequencing	Sanger sequencing
SSR1	FXAT9O003C9CR9	atcatcatcatcatc	atcatcatcatcatc
SSR2	FXAT9O003C27N8	agtttagtttagttt	agtttagtttagttt
SSR3	FXAT9O003C3EB0	ggagggagggag	ggaggaggaggaggagga ggaggaggagga
SSR4	FXAT9O003DNXS9	aaggaaaaggaaaa-ggaa	aaggaaaaggaaaaggaa
SSR5	FXAT9O003DGP7H	aggaagaagaagaag	aggaagaagaagaag
SSR6	contig00373	accaccaccaccacc	accaccaccaccacc
SSR7	contig03257	ctgctgctgctgctg	–
SSR8	contig06958	ttcttcttcttcttc	–
SSR9	contig03420	catcatcatcatcat	catcatcatcatcat
SSR10	contig01126	tattatattatatta	tattatattatatta

– No useful data obtained.

This study is the largest collection of sequencing population obtained from vegetative tissue in *Taxus* species. Results may allow for the identification of new and interesting metabolic pathways, as well as the study of molecular genetics and functional genomics. The method described here is a cost-effective technology for the identification of novel genes in medicinal plants. In addition, the sequence data described constitute an important resource for biosynthetic and biochemical research in *Taxus* species, which will lead to drug improvements in the future.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (grant No. 30873459). We thank Prof. Zhao Zhang (Peking Union Medical College, China) for her assistance in material collection.

References

1 *Liu T, Zhou Z.* Characteristics of natural Japanese yew population in Muling Nature Reserve of Heilongjiang Province, China. J For Res 2006; 17: 132–134
2 *Jennewein S, Long RM, Williams RM, Croteau R.* Cytochrome P450 taxadiene 5alpha-hydroxylase, a mechanistically unusual monooxygenase catalyzing the first oxygenation step of taxol biosynthesis. Chem Biol 2004; 11: 379–387
3 *Schoendorf A, Rithner CD, Williams RM, Croteau RB.* Molecular cloning of a cytochrome P450 taxane 10*β*-hydroxylase cDNA from *Taxus* and functional expression in yeast. Proc Natl Acad Sci USA 2001; 98: 1501–1506
4 *Jennewein S, Wildung MR, Chau M, Walker K, Croteau R.* Random sequencing of an induced *Taxus* cell cDNA library for identification of clones involved in Taxol biosynthesis. Proc Natl Acad Sci USA 2004; 101: 9149–9154
5 *Chau M, Jennewein S, Walker K, Croteau R.* Taxol biosynthesis: Molecular cloning and characterization of a cytochrome P450 taxoid 7*β*-hydroxylase. Chem Biol 2004; 11: 663–672
6 *Chau M, Croteau R.* Molecular cloning and characterization of a cytochrome P450 taxoid 2*α*-hydroxylase involved in Taxol biosynthesis. Arch Biochem Biophys 2004; 427: 48–57
7 *Jennewein S, Rithner CD, Williams RM, Croteau RB.* Taxol biosynthesis: Taxane 13*α*-hydroxylase is a cytochrome P450-dependent monooxygenase. Proc Natl Acad Sci USA 2001; 98: 13 595–13 600
8 *Croteau R, Ketchum REB, Long RM, Kaspera R, Wildung MR.* Taxol biosynthesis and molecular genetics. Phytochem Rev 2006; 5: 75–97
9 *Eisenreich W, Menhard B, Hylands PJ, Zenk MH, Bacher A.* Studies on the biosynthesis of taxol: The taxane carbon skeleton is not of mevalonoid origin. Proc Natl Acad Sci USA 1996; 93: 6431–6436
10 *Baloglu E, Kingston DGI.* The taxane diterpenoids. J Nat Prod 1999; 62: 1448–1472
11 *Sun C, Li Y, Wu Q, Luo H, Sun Y, Song J, Lui EM, Chen S. De novo* sequencing and analysis of American ginseng root transcriptome using a GS FLX Titanium platform to discover putative genes involved in ginsenoside biosynthesis. BMC Genomics 2010; 11: 262
12 *Parkinson J, Blaxter M.* Expressed sequence tags: an overview. Methods Mol Biol 2009; 533: 1–12
13 *Li Y, Luo H-M, Sun C, Song JY, Sun YZ, Wu Q, Wang N, Yao H, Steinmetz A, Chen SL.* EST analysis reveals putative genes involved in glycyrrhizin biosynthesis. BMC Genomics 2010; 11: 268
14 *Li Y, Sun C, Luo H, Li X, Niu Y, Chen S.* Transcriptome characterization for *Salvia miltiorrhiza* using 454 GS FLX. Acta Pharm Sin 2010; 4: 45
15 *Tumer NE, Clark WG, Tabor GJ, Hironaka CM, Fraley RT, Shah DM.* The genes encoding the small subunit of ribulose-1,5-bisphosphate carboxylase are expressed differentially in *Petunia* leaves. Nucleic Acids Res 1986; 14: 3325–3342
16 *Fuentes SI, Allen DJ, Ortiz-Lopez A, Hernández G.* Over-expression of cytosolic glutamine synthetase increase photosynthesis and growth at low nitrogen concentrations. J Exp Bot 2001; 52: 1071–1081

17 *Hanashiro I, Itoh K, Kuratomi Y, Yamazaki M, Igarashi T, Matsugasako J, Takeda Y.* Granule-bound starch synthase I is responsible for biosynthesis of extra-long unit chains of amylopectin in rice. Plant Cell Physiol 2008; 49: 925–933
18 *Sekhar Pagadala N, Arha M, Reddy PS, Kumar R, Sirisha VL, Prashant S, Janardhan Reddy K, Khan B, Rawal SK, Kavi Kishor PB.* Phylogenetic analysis, homology modelling, molecular dynamics and docking studies of caffeoyl-CoA–O-methyl transferase (CCoAOMT 1 and 2) isoforms isolated from subabul (*Leucaena leucocephala*). J Mol Model 2009; 15: 203–221
19 *Lefebvre S, Lawson T, Fryer M, Zakhleniuk OV, Lloyd JC, Raines CA.* Increased sedoheptulose-1,7-bisphosphatase activity in transgenic tobacco plants stimulates photosynthesis and growth from an early stage in development. Plant Physiol 2005; 138: 451–460
20 *Liu P, Goh CJ, Loh CS, Pua EC.* Differential expression and characterization of three metallothioneinlike genes in Cavendish banana (*Musa acuminata*). Physiol Plant 2002; 114: 241–250
21 *Mahmoud SS, Croteau RB.* Metabolic engineering of essential oil yield and composition in mint by altering expression of deoxyxylulose phosphate reductoisomerase and menthofuran synthase. Proc Natl Acad Sci USA 2001; 98: 8915–8920
22 *Estévez JM, Cantero A, Reindl A, Reichler S, León P.* 1-Deoxy-D-xylulose-5-phosphate synthase, a limiting enzyme for plastidic isoprenoid biosynthesis in plants. J Biol Chem 2001; 276: 22 901–22 909
23 *Kaspera R, Croteau R.* Cytochrome P450 oxygenases of Taxol biosynthesis. Phytochem Rev 2006; 5: 433–444
24 *Walker KD, Klettke K, Akiyama T, Croteau R.* Cloning, heterologous expression, and characterization of a phenylalanine aminomutase involved in Taxol biosynthesis. J Biol Chem 2004; 279: 53 947–53 954
25 *Long RM, Croteau R.* Preliminary assessment of the C13-side chain 2′-hydroxylase involved in Taxol biosynthesis. Biochem Biophys Res Commun 2005; 338: 410–417
26 *Hampel D, Mau CJD, Croteau RB.* Taxol biosynthesis: Identification and characterization of two acetyl CoA: taxoid-*O*-acetyl transferases that divert pathway flux away from Taxol production. Arch Biochem Biophys 2009; 487: 91–97
27 *Endt DV, Kijne JW, Memelink J.* Transcription factors controlling plant secondary metabolism: what regulates the regulators? Phytochemistry 2002; 61: 107–114
28 *Blackshear PJ.* Tristetraprolin and other CCCH tandem zinc-finger proteins in the regulation of mRNA turnover. Biochem Soc Trans 2002; 30: 945–952
29 *Wang L, Xu Y, Zhang C, Ma Q, Joo SH, Kim SK, Xu Z, Chong K.* OsLIC, a novel CCCH-type zinc finger protein with transcription activation, mediates rice architecture via brassinosteroids signaling. PLoS One 2008; 3: e3521
30 *Armstead IP, Skøt L, Turner LB, Skøt K, Donnison IS, Humphreys MO, King IP.* Identification of perennial ryegrass (*Lolium perenne* [L.]) and meadow fescue (*Festuca pratensis* [Huds.]) candidate orthologous sequences to the rice *Hd1* (*Se1*) and barley *HvCO1 CONSTANS*-like genes through comparative mapping and microsynteny. New Phytol 2005; 167: 239–347
31 *Lijavetzky D, Carbonero P, Vicente-Carbajosa J.* Genome-wide comparative phylogenetic analysis of the rice and *Arabidopsis* Dof gene families. BMC Evol Biol 2003; 3: 17
32 *Vij S, Tyagi AK.* A20/AN1 zinc-finger domain-containing proteins in plants and animals represent common elements in stress response. Funct Integr Genomics 2008; 8: 301–307
33 *Zhang J, Peng Y, Guo Z.* Constitutive expression of pathogen-inducible OsWRKY31 enhances disease resistance and affects root growth and auxin response in transgenic rice plants. Cell Res 2008; 18: 508–521
34 *Chagné D, Chaumeil P, Ramboer A, Collada C, Guevara A, Cervera MT, Vendramin GG, Garcia V, Frigerio JM, Echt C, Richardson T, Plomion C.* Cross-species transferability and mapping of genomic and cDNA SSRs in pines. Theor Appl Genet 2004; 109: 1204–1214