The *Taxus* genome provides insights into paclitaxel biosynthesis

Xingyao Xiong^{1,2,9}, Junbo Gou^{2,9}, Qinggang Liao^{2,9}, Yanlin Li^{1,3,9}, Qian Zhou^{2,4}, Guiqi Bi²,
Chong Li², Ran Du², Xiaotong Wang², Tianshu Sun², Lvjun Guo⁵, Haifei Liang², Pengjun Lu²,
Yaoyao Wu², Zhonghua Zhang⁶, Dae-Kyun Ro^{2,7}, Yi Shang⁸, Sanwen Huang^{2*}, Jianbin Yan^{2*}

¹College of Horticulture, Hunan Agricultural University, Changsha 410128, China.

²Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Shenzhen Key Laboratory of Agricultural Synthetic Biology, Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518124, China.

³Engineering Research Center for Horticultural Crop Germplasm Creation and New Variety Breeding, Ministry of Education, Changsha 410128, China.

⁴Peng Cheng Laboratory Artificial Intelligence Research Center No.2, Xingke 1st Street, Nanshan, Shenzhen, 518055, China.

⁵MOE Key Laboratory of Bioinformatics, Tsinghua-Peking Joint Center for Life Sciences, School of Life Sciences, Tsinghua University, Beijing 100084, China.

⁶College of Horticulture, Qingdao Agricultural University, Qingdao 266109, China.

⁷Department of Biological Sciences, University of Calgary, Calgary, AB, T2N1N4, Canada.

⁸The AGISCAAS-YNNU Joint Academy of Potato Sciences, Yunnan Normal University, Kunming, China.

*Correspondence to: jianbinlab@caas.cn (J.Y.); huangsanwen@caas.cn (S.H.)

⁹These authors contributed equally to this work.

Abstract

The ancient gymnosperm genus *Taxus* is the exclusive source of the anticancer drug paclitaxel, yet no reference genome sequences are available for comprehensively elucidating the paclitaxel biosynthesis pathway. We have completed a chromosome-level genome of *Taxus chinensis* var. *mairei* with a total length of 10.23 Gb. *Taxus* shared an ancestral whole-genome duplication with the coniferophyte lineage and underwent distinct transposon evolution. We discovered a unique physical and functional grouping of *CYP725As* (cytochrome P450) in the *Taxus* genome for paclitaxel biosynthesis. We also identified a gene cluster in the taxadiene biosynthesis, which was mainly formed by gene duplications. This study will facilitate the elucidation of paclitaxel biosynthesis and unleash the biotechnological potential of *Taxus*.

One Sentence Summary

A chromosome-level genome assembly of *Taxus chinensis* var. *mairei* uncovers its unique genome evolution process and genetic architectures for the paclitaxel biosynthesis pathway.

Keywords

Taxol, Cytochrome P450, Gymnosperm, Jasmonate, Gene cluster

Main

Taxaceae (*Taxus spp.*), a widespread nonflowering conifer with substantial economic value, contains six extant genera and over 28 species¹. *Taxus* is the largest genus in Taxaceae, including common species such as *T. chinensis*, *T. brevifolia*, and *T. baccata*, and it is mainly distributed in Asia, North America, and Europe². For decades, *Taxus* has served as a natural source to obtain paclitaxel (trade name Taxol), a well-known chemotherapy agent against various cancers³. Nevertheless, plant-derived paclitaxel suffers from a short supply due to its low abundance in *Taxus*, limiting its clinical application. In this instance, multiple strategies have been employed to address supply issues⁴, and promising progress has been made in chemical⁵ and semichemical synthesis⁶, direct extraction from *Taxus* cell lines⁷, fermentation of endophytic paclitaxel producing fungi⁸, and metabolic engineering of paclitaxel production using heterologous systems⁹.

As a tetracyclic diterpene, paclitaxel is biosynthesized by a complex metabolic pathway¹⁰. The paclitaxel pathway starts with geranylgeranyl diphosphate (GGPP) synthesis through condensation of isoprenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP)¹¹. GGPP is then cyclized by taxadiene synthetase (TS), generating a unique diterpene skeleton, taxadiene¹². Taxadiene is subsequently decorated by a series of reactions including hydroxylation, oxidation, epoxidation, acylation and benzylation, to generate the final product via catalysis by various enzymes (e.g., hydroxylase, oxidase, epoxidase, oxomutase, and transferase)¹³⁻¹⁵. To date, over twenty enzymes have been identified in the paclitaxel biosynthetic pathway. However, several essential steps such as C1 hydroxylation, C9 oxygenation and oxetane formation in the paclitaxel

biosynthetic pathway, remain to be clarified. Moreover, studies have shown that jasmonates (JAs), gibberellin, auxin, and ethylene are involved in the regulation of paclitaxel biosynthesis to maintain a delicate balance between growth and defense in *Taxus*¹⁶⁻¹⁸. Several transcription factors (TFs), including AP2/ERF, WRKY, MYC, and MYB, have been found to regulate the expression of paclitaxel biosynthetic genes¹⁹⁻²¹. However, the comprehensive regulatory mechanisms underlying the growth-defense tradeoff are still poorly understood.

Although a complete *Taxus* genome sequence can provide valuable bioinformatic and genetic resources to understand paclitaxel biosynthesis and regulatory mechanisms in depth, the sizeable and complex genome (2C-value: 22.3-24.3 picogram) of *Taxus* has hindered its *de novo* draft genome assembly to date²². Here, we successfully assembled the *Taxus* genome, and presented a reference-grade genome sequence of *T. chinensis* var. *mairei* containing 10.23 Gb data with contig N50 of 2.44 Mb, 9.86 Gb of which was assigned to 12 pseudochromosomes. We demonstrated that the CYP725A (cytochrome P450) genes, closely related to paclitaxel biosynthesis, have evolved independently in a unique physical and functional grouping in the *Taxus* genome. Moreover, we uncovered a gene cluster for taxadiene biosynthesis that contains a new type of taxadiene synthase. These results contribute to our understanding of the biological and evolutionary questions regarding paclitaxel biosynthesis and provide insights into the genome structure and organization of gymnosperms.

Results

Taxus genome sequencing, assembly and annotation

To build a chromosome-level genome assembly of *Taxus*, genomic DNA was extracted from endosperm calli. The endosperm of *T. chinensis* var. *mairei* seeds with haploid chromosomes was used to culture the callus, as it could prevent the influence of heterozygous elements in the genome assembly. *K*-mer analysis showed that the genome size of *T. chinensis* var. *mairei* was approximately 10 Gb (Supplementary Fig. 1), which is consistent with the results from the flow cytometry tests²³. A *de novo* assembly of the *Taxus* genome was achieved by PacBio continuous long reads (318.05 Gb) and augmented with Illumina whole-genome sequencing (WGS) reads (693.73 Gb) (Supplementary Table 1). After the application of high-throughput/resolution chromosome conformation capture (Hi-C) (Supplementary Table 2), 9.86 Gb of sequence data could be assigned to 12 pseudochromosomes (Supplementary Fig. 2, Supplementary Table 3), which covered 96.28% of the genome (Supplementary Table 4). We finally obtained the genome sequence with a total length of 10.23 Gb and a contig N50 of 2.44 Mb (Fig. 1a, Supplementary Table 4).

Based on the genomic information, 42,746 protein-coding genes were further annotated by integrating transcriptome data, homologous alignments, and *ab initio* gene models. 73.02% of the genes (31,214 out of 42,746) could be supported by RNA-seq data (Supplementary Fig. 3, Supplementary Table 5). The BUSCO analysis further demonstrated that 1,052 of the 1,614 core

genes were complete, showing relatively high completeness of the assembled genome in gymnosperms (Supplementary Table 6). Furthermore, 36,518 coding genes, accounting for 85.43% of the total predicted genes, were assigned to functional categories with an *E*-value less than 10^{-5} (Supplementary Table 7).

Taxus experienced a whole-genome duplication event in the cupressophyte clade

Given that the whole genome duplication (WGD) is a significant evolutionary force contributing to the expansion of plant genome size²⁴, we investigated whether *Taxus* had experienced any WGD events. We built a paralogous gene pair set by performing an all-against-all blastp search. The number of synonymous substitutions per synonymous site (*Ks*) of paralogs was calculated using the gene pair set. As shown in Fig. 1b, the number of *Ks* frequency exhibited an apparent decay without a natural distribution with increasing of *Ks* values, which indicated that no recent WGD event occurred in the *Taxus* genome. Moreover, we noticed that most of the *Ks* values were less than 0.8, indicating genes duplication in combination with saturation and stochasticity effects may obscure WGD²⁴. We further used MCScanX to produce 8,148 syntenic gene pairs from the all-against-all blastp data and entered them into the *Ks* and distance-transversion rate at fourfold degenerate sites (4DTv) calculations. The results showed two signature peaks located at 2.1 for *Ks* (Fig. 1b) and 0.7 for 4DTv (Supplementary Fig. 4), suggesting the presence of an ancient WGD in *Taxus*. Together with previous studies that revealed an ancient WGD event (WGD- ζ) in the common ancestor of angiosperms and gymnosperms^{24,25}, all the results mentioned above suggested that *Taxus* shared the common ancient WGD with other coniferophyte lineages.

Taxus genome expansion is linked with retrotransposons

Except for the role of WGD in enlarging the *Taxus* genome size, we noticed that repetitive sequences constituted a significant component of the *Taxus* genome (Supplementary Table 8). There was a total of 7.79 Gb of repetitive sequences, occupying 76.09% of the entire genome (Supplementary Table 8). Among these repetitive sequences, long terminal repeat (LTR) retrotransposons accounted for the highest proportion at 52.38% (Supplementary Table 8). The insertion time analysis revealed that LTR insertion was a continuous process, and approximately 40% of the insertions occurred 8 to 24 million years ago (MYA) (Fig. 1c). This feature of continuous insertion in the *Taxus* genome was significantly different from that in the rice genome, where almost 95% of LTR insertions occurred within the last 5 million years²⁶. Considering that LTR insertion in Norway spruce and ginkgo mainly occurred 12-24 and 16-24 MYA^{23,25}, the continuous insertion of LTRs might be a common phenomenon in gymnosperms.

To further explore the evolution of LTR in *Taxus*, we analyzed the phylogenies of LTR retrotransposons in a few representative gymnosperm and angiosperm plants. Amino acid sequence similarities within the reverse transcriptase domain of the Ty3/Gypsy retrotransposons (Gypsy) and Ty1/Copia retrotransposons (Copia) were used to construct phylogenetic trees. As shown in Fig. 1d, the Gypsy superfamily members of the gymnosperms ginkgo and picea were distributed in families II-VII, while those of angiosperms mainly belonged to family VIII. In contrast, *Taxus* Gypsy elements were not only distributed in families II to VIII but also evolved a highly species-specific family (family I), suggesting the expansion of specific Gypsy elements

after *Taxus* speciation. Similarly, the unique expansion phenomenon in *Taxus* was also observed in the phylogenies of the Copia superfamily (Fig. 1d). Moreover, family V consisted of only *Taxus* LTRs in the Copia phylogenetic tree displaying a *Taxus*-specific amplification burst. In addition, *Taxus* was distributed in family IV, where the gymnosperms ginkgo and picea were located, and families I-III also contained angiosperms, suggesting that *Taxus* LTRs were placed in a unique position compared to other selected species. These results suggested that the Gypsy and Copia superfamilies of *Taxus* have undergone a relatively unique evolutionary pattern, especially the specific Gypsy family I and Copia family V.

Evolution of gene families and elevated secondary metabolism in *Taxus*

To understand the context of metabolic networks during *Taxus* evolution, we compared orthologous genes between *Taxus* and selected gymnosperms, angiosperms and cryptogams (Fig. 1e). In the 35,298 identified orthologous gene families (Supplementary Table 9), we found that 6,533 gene families were shared by the selected species, illustrating their evolutionary conservation (Fig. 1e). Compared with the selected species, 2,339 gene families were exclusive to *Taxus* (Fig. 1e). Besides, 1,378 gene families experienced loss, while 142 and 41 families underwent expansion and contraction in the *Taxus*, respectively (Fig. 1f).

Taxus contains 9,747 unique genes (Fig. 1e, Supplementary Table 10), many of which are enriched in the biosynthesis of specialized metabolites, including terpenes, phenylpropanoids, and flavones (Supplementary Table 11). For instance, 57 gene families were annotated to be

cytochrome P450 (CYP450) gene families (Supplementary Table 10). Gene expansion analysis demonstrated that 979 genes were enriched in ADP binding, oxidoreductase activity, flavin adenine dinucleotide binding, transferase activity, and signal transduction, etc. (Supplementary Fig. 5, Supplementary Table 12), with eight gene families being associated with CYP450 (Supplementary Table 13). Pfam functional analysis further showed that the *Taxus* genes were enriched in CYP450 gene families (PF00067.22, $p < 0.01$) and TFs (PF13837.6, $p < 0.01$ and PF00847.20, $p < 0.01$) (Supplementary Table 14). KEGG analysis indicated that the gained and expanded gene families were enriched in a total of 36 and 41 KEGG pathways, respectively, including one phenylpropanoid (ko00940) and three terpenoid metabolic pathways (ko00900, ko00130, ko00902) (Supplementary Tables 11 and 15).

Evolution and genomic organization of *Taxus* CYP450s

Given that CYP450s participate in almost half of the enzymatic reactions in paclitaxel biosynthesis²⁷, we analyzed *Taxus* CYP450 families and identified 649 CYP450 genes from the present genome using the reported HMM model (PF00067). These CYP450s can be divided into two catalogs (A-type and non-A-type). The A-type CYP450s included only the CYP71 clan, which consisted of 17 families and 325 genes (Supplementary Fig. 6 and Supplementary Table 16), while the non-A-type CYP450s contained 12 clans that were composed of 27 families and 324 genes (Supplementary Fig. 7 and Supplementary Table 16). Phylogenomic analyses showed that the CYP750 and CYP725 families were significantly expanded in *Taxus* compared to 68 other representative species, which covered Zygnematophyceae and Sapindaceae (Fig. 2a, Supplementary Figs. 8 and 9, and Supplementary Table 17). The CYP750 family was reported to

participate in the biosynthesis of thujone monoterpene, which is involved in defense responses (e.g., resistance against herbivore feeding)²⁸, while CYP725 genes were known to contribute to paclitaxel biosynthesis²⁹. Phylogenetic analysis of these CYP725 genes further showed that they could be categorized into the CYP725A and CYP725B subfamilies (Fig. 2b). The CYP725A subfamily (a total of 80 genes) exhibited specificity to *Taxus*, whereas the CYP725B subfamily was universal in gymnosperm plants (including *Picea*, *Cycas*, *Ginkgo*, and *Taxus*) (Fig. 2b), which suggested that CYP725A underwent independent evolution in *Taxus*. Considering that all the previously defined CYP450 genes in the paclitaxel pathway belong to the CYP725A subfamily, these results suggested that the expansion of the CYP725A subfamily played vital roles in the evolution of paclitaxel biosynthesis in *Taxus*.

We noticed that most CYP725A genes (64%) were located on pseudochromosome 9 (Fig. 2c and 2d), exhibiting a distinct aggregated distribution. Gene aggregation analysis further revealed that the *Taxus* CYP450 genes were not distributed randomly but trended to organize into different gene groups, 24 of which were detected in the genome (Fig. 2c). We found that nearly all these groups, except groups 1.2 and 5.1, contained gene members from no more than three CYP450 families, and ten groups had only one CYP450 family (Supplementary Table 18), suggesting that the grouping of CYP450 genes on the genome had an obvious family aggregation pattern. Furthermore, as an essential phytohormone in the biosynthesis of secondary metabolites³⁰, jasmonate is closely related to the expression regulation of CYP450 genes in *Taxus* (Fig. 2e). Under jasmonate treatment, 8 groups showed an increased expression level, 12 groups showed the inhibition of gene expression (Fig. 2e, Supplementary Fig. 10 and Supplementary Table 18). These results suggested a tendency that the CYP450s in the majority of groups were coexpressed

under jasmonate treatment, implying that the grouping of CYP450 genes had some coordination of physiological functions.

The gene expression levels of four gene groups (group 9.1-9.4) on pseudochromosome 9 were upregulated most significantly in the presence of jasmonate (Fig. 2e). More interestingly, groups 9.1 and 9.2 contained all known CYP725A subfamily genes related to paclitaxel biosynthesis and 12 undefined CYP725As (Fig. 2f). The expression profiles of these two groups of CYP725A genes showed that 88% of CYP725As were highly expressed in roots, 79% of CYP725As were highly expressed in the high paclitaxel-yielding cell line (HC), and 88% of CYP725As were upregulated after jasmonate treatment (Supplementary Table 19), which is consistent with the results on the increased level of baccatin III and paclitaxel (by 3-5 times) in the *Taxus* cell line under jasmonate treatment (Supplementary Fig. 11). These results suggested that the two groups were likely to contain most of the paclitaxel pathway genes that arose during *Taxus* evolution.

Taxadiene biosynthetic genes are arranged in gene clusters

PlantSMASH³¹ analysis further showed that a gene cluster related to terpene biosynthesis was presented in group 9.2 (Fig. 3a and Supplementary Table 20). The gene cluster contained two TS genes (*TS2* and *TS3*, sharing 99.96% nucleotide sequence identity), two *T5αH* genes (*T5αH1* and *T5αH2*, sharing 98.67% nucleotide sequence identity), and two novel CYP725As (Fig. 3a, Supplementary Table 21, and Supplementary Figs. 12 and 13). Moreover, the genes in the cluster showed a highly coordinated tissue expression pattern and expression

consistency in response to jasmonate treatment (Fig. 3a), suggesting that the genes could be functionally related. *TS2* and *TS3* were located adjacent to *T5aH1* and *T5aH2* (Fig. 3a), suggesting that the genes involved in the first two paclitaxel biosynthetic steps were organized by a tandem gene duplication event during *Taxus* genomic evolution. The K_s value of these duplicated genes suggested that this TS-*T5aH* duplication occurred at approximately 1.15 MYA. In addition to the *TS* and *T5aH* genes assembled in the cluster, additional *TS* (*TS1*) and *T5aH* (*T5aH3*) genes are located downstream and upstream of the cluster, respectively (Fig. 3a, Supplementary Figs. 12 and 13). Biochemical assays further confirmed that *TS1/2* and *T5aH1/2/3* have taxadiene synthase activity (Fig. 3b, Supplementary Figs. 14) and taxadiene-4(5),11(12)-diene-5 α -hydroxylase activity (Fig. 3c, Supplementary Figs. 14 and), respectively, demonstrating that the copied genes possessed the corresponding enzyme activities in *T. chinensis* var. *mairei*.

We further studied the kinetic properties of *TS1* and *TS2*. The K_m value of *TS2* was approximately 1.5 times higher than that of *TS1*, but the turnover number (k_{cat}) of *TS2* was approximately 3 times greater than that of *TS1*, indicating that *TS2* has a higher catalytic efficiency than *TS1* (Fig. 3d). Moreover, exogenous jasmonate treatment resulted in an obviously higher level of *TS2* than *TS1* transcripts (Fig. 3e), implying that *TS2* could play a role in paclitaxel biosynthesis in response to different environmental and developmental cues via JA signaling. Sequence identity analysis showed that *TS2* shared only 77-78% protein sequence identity with *TS1* and *T. brevifolia* *TS* (TbTS), which is much lower than the sequence similarity (over 90%) within the previously reported *TS* genes (Supplementary Table 21), suggesting that *TS2* is a unique *TS* gene diverged from *TS1* and TbTS. Phylogenetic tree analysis further

confirmed that a *Taxus*-specific gene duplication event at approximately 33.2 MYA resulted in two distinct types of TS genes (Supplementary Fig. 15), demonstrating that taxadiene synthases were encoded by two types of TS genes resulting from gene duplication events in *Taxus*. Together, these results suggested that the genes involved in the two initial steps of the paclitaxel biosynthesis pathway were arranged in a gene cluster named the “taxadiene gene cluster”. The taxadiene gene cluster might be formed by gene duplications and neofunctionalization in *Taxus* and somewhat similar to previous studies on operon-like gene clusters in plants^{32,33}.

Furthermore, we established a gene-to-gene coregulation network using three rounds of subtraction screening with RNA-Seq datasets. The network could cover all known paclitaxel biosynthetic genes (Supplementary Figs. 16 and 17, Supplementary Tables 22 and 23), indicating its comprehensiveness and high credibility. We identified 17 CYP725A genes, 3 transferases, and 10 TFs with this network, which was strongly associated with known paclitaxel biosynthetic genes (Supplementary Table 24). Real-time quantitative PCR assays confirmed that the expression of certain genes could be induced by jasmonates (Supplementary Fig. 18), implying that their encoded proteins could be investigated as potential enzymes of paclitaxel biosynthesis. Together, these results outlined the biosynthesis pathway of paclitaxel in *T. chinensis* var. *mairei* (Fig. 3f) and provided valuable genetic resources for improving paclitaxel production through genetic breeding and synthetic biology.

Discussion

The absence of a chromosome-level genome sequence from *Taxus* has prevented in-depth phylogenomic studies of *Taxus*. Our study provides an example assembly of a complex genome in trees using various sequencing technologies on DNA from endosperm calli containing haploid chromosomes. Flow cytometry analysis indicated that the nuclear genome (2n) size of the diploid cells of *Taxus* was approximately 20.80 ~ 24.85 Gb, nearly twice the haploid genome size evaluated by *k*-mer analysis (Supplementary Fig. 1). The vast majority of HiFi sequences from *Taxus* leaves (diploid) could be mapped to the haploid genome, up to 95%. Moreover, 75.81% of the 228,762,501 SNPs were, 44.97% of the 847,935 InDels were, and 85.64% of the 64,927 structural variants were heterozygous. Taken together with the low heterozygosity (0.02%) of *T. chinensis* var. *mairei*, these results demonstrated that the haploid genome assembly could basically represent its diploid genome, showing the advantages of using endosperm calli for genome assembly.

We found that the complete BUSCOs only increased from 64.7% to 65.2% when the N50 value was increased from 637 kb to 2.44 Mb during the genome assembly. The low BUSCO value might be due to the limitations of the BUSCO reference dataset. The latest dataset version is embryophyta_odb10 (2020-09-10), containing 1,614 genes from single-copy genes of 50 species, including two bryophytes (*Physcomitrella patens* and *Marchantia polymorpha*), one fern (*Selaginella moellendorffii*), and 47 seed plants (all are angiosperms), but has not included any genes of gymnosperms. Consistently, across all of the reported gymnosperm genomes, except for *Gnetum montanum*, the BUSCO values were not higher than 73%, and four of these genomes had values lower than 51% (Supplemental Table 6). The BUSCO value of the *Taxus* genome was 65.2%, similar to that of *Ginkgo biloba* (69.4%) and *Pseudotsuga menziesii* (67.8%). To

assess the *Taxus* genome quality more comprehensively, we mapped the Illumina DNA sequencing data (~ 693 Gb) for the genome survey onto the assembled genome and found that up to 99.60% of the sequencing data could be mapped, indicating the integrity of the genome assembly. Moreover, we collected transcriptome data from *Taxus* organs, comprehensively covering eight tissues and cell lines (root, stem, leaf, bark, male strobili, female strobilus, high-yield paclitaxel cell lines, and low-yield paclitaxel cell lines), and mapped the sequencing data to the *Taxus* genome. The results showed that the average overall mapping rate of transcriptome data to the genome reached 90.45% (Supplementary Table 25), suggesting the integrity of functional genes in the genome.

The *Taxus* genome contains 4.08 Gb LTR retrotransposons, including 87.28% Gypsy and 12.35% Copia retrotransposons, and a small proportion of unknown LTRs (0.37%) (Supplementary Table 8). The LTR distribution analysis showed that LTRs tended to be distributed throughout the entire chromosome (Supplementary Fig. 19). In particular, Copia tends to be enriched at the two ends of the chromosomes, while Gypsy is more enriched at the chromosome ends and central areas. Compared with previous studies in groundnuts³⁴, the two genomes exhibited obvious differences in LTR distribution. The LTR retrotransposons of the groundnut genome are mainly distributed in the central region of chromosomes, close to the centromere. This difference may come from the large disparity in genome size and the difference between angiosperms and gymnosperms.

The LTR insertions in the *Taxus* genome mainly occurred 8 to 24 MYA during the long insertion period (4-60 MYA) (Fig.1c and Supplementary Fig. 20), while the primary insertion time of

LTRs in spruce and ginkgo was 12-24 and 16-24 MYA within their insertion span from 4 to 64 MYA^{25,35}. These results suggested that the *Taxus* genome has a similar LTR insertion time trend to that in the spruce and ginkgo genomes. The very long insertion time phenomenon might be related to the evolutionary characteristics of gymnosperms. It is generally accepted that gymnosperms belong to slow-evolving plants. Their morphology is highly conserved, which is supported by the high similarity between extant species and fossil records. Previous studies have shown that angiosperms and gymnosperms differ considerably in their mutation rates of molecular evolution per unit time, with gymnosperm rates being, on average, seven times lower than those of angiosperm species³⁶. For this reason, an insertion time longer than 60 million years is common in gymnosperm genomes because of the much lower mutation rate. For example, up to 8.27% of LTRs were inserted into the ginkgo genome over 60 million years, and 13.31% of LTRs inserted into the Spruce genome over 60 million years^{25,35}.

In addition to CYP450 enzymes, acetyltransferases also play an essential role in paclitaxel biosynthesis. We found 127 BAHD acyltransferases by identifying their conserved motifs (HXXXD and DFGWG) (Supplementary Fig. 21). The BAHD acyltransferases in the *Taxus* were mainly distributed in Clades I, II, VI, and V. Clade V can be divided into three groups (Groups I-III), among which Group I contains all known BAHD acyltransferases in the paclitaxel biosynthesis pathway. It would be worthwhile to investigate whether group I contains other acyltransferases that function in the paclitaxel biosynthesis in the future (Supplementary Table 26). PlantSMASH analysis indicated that the acyltransferase genes are not organized into any gene clusters. Genomic location analysis showed that the BAHD acyltransferase genes in paclitaxel biosynthesis were mainly distributed on chromosomes 1 and 9 (Supplementary Fig.

17). Furthermore, TAT2 was colocalized with CYP450s in gene group 9.2 (Fig. 2c and Supplementary Fig. 17). The relationship between CYP725As and acetyltransferases in paclitaxel evolution is an interesting aspect to study in the future.

In the *Taxus* genome, a total of 34 gene clusters related to secondary metabolism were found, including 13 saccharides, seven terpenes, one alkaloid, one saccharide-terpene, one saccharide-polyketide, one lignan-terpene, one terpene-alkaloid, and nine putative gene clusters (Supplemental Tables 20 and 27). Two gene clusters (clusters I and II) belonging to the terpene cluster were involved in paclitaxel biosynthesis because cluster I contained the TS2, TS3, T5 α H2, and T5 α H3 genes, and cluster II included TS1. Except for these five genes, other related enzymes in the paclitaxel synthesis pathway were not included in any gene clusters. However, we found that most of the known genes involved in paclitaxel biosynthesis, including TAT2, DBAT, TS1/2/3, T7 β H1/2, T13 α H1/2, T10 β H1/2/3, T5 α H1/2/3, and T14 β H, are located on a small 71.82 Mb region on chromosome 9 (designated T13 α H2-DBAT segment: 19994572-91,811,351 bp, Supplemental Fig. 22). Therefore, many genes that play roles in different steps of the paclitaxel synthesis pathway are located in a limited genomic region, implying that there might be a certain coordinated regulatory mechanism of their gene expression. It would be an important future project to investigate whether the genes are organized in a larger-scale gene cluster to achieve better collaborative expression.

To date, all known TS enzymes are homologous to TS1 (amino acid homology > 90%) (Supplementary Table 21A). Our study demonstrated that taxadiene synthases could be encoded

by two distinct types of TSs resulting from gene duplication events in *Taxus*. As a representative of the new type of TS enzyme, TS2 only has approximately 77-78% amino acid homology with the reported TS enzymes (Supplementary Table 21). It exhibits a higher catalytic efficiency than TS1 (Fig. 3d) and more robust induced expression characteristics in treatment with jasmonates (Fig. 3e). The different properties of these two types of TS enzymes implied a new *Taxus* defense regulation mechanism. In *Taxus*, the excessive synthesis of taxanes is not conducive to its growth or development although chemicals play an essential role in defense responses. Therefore, it is necessary to accurately and efficiently control the taxane level in cells in response to environmental changes. Our results provide a new hypothesis to explain the regulation of taxane levels in plant cells. When there are no biotic or abiotic stresses, jasmonate signaling is blocked, and TS1 is responsible for taxane biosynthesis to maintain taxanes at a basic level. However, once insect attack or other stresses occur, jasmonate signaling will be activated, and TS2 is rapidly expressed to quickly increase the taxane content in cells.

In addition, we tried to explore the application potential of TS2 in bioengineering. Bian et al. reported an engineered *E.coli* strain with TbTS (belonging to Type I) for the taxadiene product³⁷. We replaced the TbTS gene with the TS2 gene (Supplementary Fig. 23). After 60 hours of fermentation, we were surprised to find that the taxadiene titer from the strain containing TS2 was over ten times higher than that containing TbTS, while the OD₆₀₀ of the two strains was not much different (Supplementary Fig. 23). The result showed the great potential of TS2 in bioengineering to produce taxadiene in the future.

We also explored the function of two unknown CYP725As (55305455 and 55326109) in the taxadiene cluster using the well-established T5 α H reaction assay. We incubated yeast microsomes that included 55305455, T5 α H1, and CPR with the taxadiene as substrate at the same time and analyzed the reaction products by GC-MS. As shown in Supplemental Figure 24, we only detected OCT, iso-OCT, and taxa-4(20),11(12)-dien-5 α -ol, which can be obtained by catalyzing taxadiene by the T5 α H1 enzyme. The same result was obtained with 55326109 protein in the reaction system. These results suggested that the unknown CYP725s are not involved in the subsequent reaction catalyzed by T5 α H. However, the tissue expression specificity of 55305455 and 55326109 was similar to that of TS and T5 α H in the cluster, and both of them exhibited higher expression levels in roots than in leaf and bark tissues (Fig. 3a). The real-time PCR assay validated that their expression was induced by jasmonate in *Taxus* cells (Supplementary Fig. 18), which is consistent with paclitaxel accumulation (Supplemental Fig. 11). Moreover, the updated gene-to-gene coregulation network showed that 55305455 and 55326109 were correlated with DBAT and T5 α H1, respectively (Supplemental Fig. 16). These results indicated that the two CYPs may play a role in paclitaxel biosynthesis and metabolism, and are worthy of in-depth study in the future.

Methods

Plant materials and genome sequencing

Seeds of a single female *T. chinensis* var. *mairei* were collected from the natural range of *Taxus* (GPS: N 113° 89' 55", E 28° 26' 32") in the Liuyang region, Changsha city, Hunan Province, China, in November 2015. Single embryos and endosperm were induced as calli^{23,38}.

For sequencing of the haploid tissue, DNA was extracted from the endosperm callus of *T. chinensis* var. *mairei*²³. The DNA quality was checked by agarose gel electrophoresis and a Qubit fluorimeter (Thermo Fisher, Inc.). The paired-end libraries with a 500 bp insert length were prepared by following the Illumina protocols. Sequencing of the library was performed on the Illumina HiSeq 2500 system. For the PacBio Sequel analysis, SMRTbell™ libraries were prepared according to the manufacturer's protocol of the sequencing platform. Four independent Hi-C libraries were constructed and sequenced on an Illumina HiSeq 2500 (PE125 bp) at Annoroad Gene Technology Co., Ltd. (Beijing, China).

For circular consensus sequencing (CCS), genomic DNA was isolated from plant leaves using the cetyltriethyl-ammonium bromide method. A 15 kb DNA SMRTbell library was constructed and sequenced on a Pacbio Sequel II platform, the sequencing reads known as highly accurate long reads, or HiFi reads.

Genome assembly and gene annotation

The uncorrected PacBio reads were assembled using wtdbg2³⁹, the fastest sequence assembler for long noisy reads. The assembly reached the best continuity with the following parameters: -k 0 -p 19 -K 5000 -S1 --aln-noskip --tidy-reads 5000 --edge-min 2 --rescue-low-cov-edges. The software Arrow in the GenomicConsensus package (<https://github.com/Pacific>

Biosciences/GenomicConsensus) was applied to generate the consensus sequences from the primary assembly. The raw PacBio reads were aligned to the assembly of red bean³⁹ using pbalign (v0.3.1) with default parameters, and then the alignment was passed to Arrow (v2.2.2) to produce the corrected assembly. The consensus process was performed iteratively twice. Further polishing of the assembly genome was conducted using Pilon⁴⁰ with Illumina data, with the following parameters: --fix all --mindepth 0.4 --K 65 --threads 24 --minmq 30 --minqual 30 --changes.

For Hi-C assembly, the clean Hi-C sequencing data were mapped to the genome draft by HiC-Pro (V2.7.8)⁴¹, and the library quality was assessed by counting the number of uniquely valid paired-end reads. Only unique valid paired-end reads were maintained for downstream analysis. We used Hi-C data to align and correct the contigs for misassembly through Juicer⁴² pipeline and 3D-DNA pipeline⁴³. The assembly package, Lachesis⁴⁴ was applied to perform clustering, ordering and orienting based on the normalized Hi-C interactions. For each pseudochromosome group, the exact contig order and directions were obtained through a weighted directed acyclic graph (WDAG). We filled the gaps among contigs in the pseudochromosomes using TGS-Gapcloser (v1.01)⁴⁵ by two rounds with CLR and HiFi data (26 Gb), respectively. After the filling progress, we further removed the redundant contigs that were not anchored to the chromosomes by Purge Haplotigs (v1.03)⁴⁶.

For assembly assessment, the RNA-seq reads of eight tissues, including mfs, mfl, mfb, mfr, mtf, mtl, mtb and mtr, and HC and LC were mapped to assess the assembly quality (mfs: female strobilus, mfl: female leaf, mfb: female bark of stem, mfr: female root; mtf: male strobili, mtl: male leaf, mtb: male bark of stem, and mtr: male root). The average mapping rate was calculated with the HISAT2⁴⁷ with parameters --score-min L, 0, -0.1.

For repeat annotation and analyses, repetitive elements in the *Taxus* genome were identified through a combination of *de novo* and homology-based approaches. *De novo* prediction of repeat elements (RE) was carried out using RepeatModeler (v1.0.1, <http://www.repeatmasker.org/RepeatModeler/>). For homology-based annotation, the RE libraries from Repbase⁴⁸, TIGR (The Institute for Genomic Research)⁴⁹ and the annotated *Ginkgo biloba* genome were merged with the *de novo*-derived library to create the whole dataset. The dataset was then used to mask identified TEs in the *Taxus* genome with RepeatMasker (v4.0.5, <http://www.repeatmasker.org>). We identified LTRs by referring LTR_retriever method⁵⁰. Specifically, LTR_finder⁵⁰ and LTRharvest⁵¹ were first used to identify all the existing LTR sequences in the *Taxus* genome according to the basic sequence rules of LTRs. The candidate LTR RTs were filtered to remove non-LTR RT repeat elements or those with large amounts of tandem repeats or gaps. Especially in fragmented genome assemblies, these requirements hugely reduce the number of LTR RT candidates but ensure that only full-length LTR RTs are analyzed. We integrated the results and discarded false positives by the LTR_retriever pipeline and then estimated insertion times based on $T=D/2\mu$, where D is the divergence rate and μ is the neutral mutation rate (7.34573×10^{-10})³⁶.

For the annotation of protein-coding genes, gene structure prediction was performed using *ab initio*, homology-based and RNA-seq-based pipelines. For the *ab initio* annotation, SNAP⁵², Augustus⁵³ and GlimmerHMM were applied. Eight species, *Arabidopsis thaliana*⁵⁴, *Oryza sativa*⁵⁵, *Gnetum montanum*⁵⁶, *Picea abies*²⁵, *Ginkgo biloba*³⁵, *Selaginella moellendorffii*⁵⁷, *Pinus taeda*⁵⁸ and *Amborella trichopoda*⁵⁹ were chosen for homology annotation to predict protein-coding genes by GeneWise⁶⁰. To generate annotation results based on transcripts, RNA-seq alignment files were generated using TopHat2⁶¹ and assembled by Cufflinks⁶², and the program PASA⁶³ was used to align spliced transcripts and annotate candidate genes. Finally, gene models

predicted from three approaches were merged by EVM⁶⁴. The functions of protein-coding genes were identified by mapping sequences against the Gene Ontology⁶⁵, InterProScan⁶⁶, Swiss-Prot (<http://www.uniprot.org/>)⁶⁷, EMBL database⁶⁸, and TAIR databases⁶⁹.

Identification of whole genome duplication

Genome-wide duplications were searched in the *Taxus* genome. Self-alignment of the assembled genome sequence was performed using metablast as described previously⁷⁰. All-vs-all paralog analysis in the *Taxus* genome was performed using reciprocal best hits from primary protein sequences by self-Blastp in *Taxus*. Reciprocal best hits are defined as reciprocal best Blastp matches with an *E*-value threshold 1e-5, c-score (Blast score/best Blast score) threshold of 0.3⁷¹, and alignment length threshold of 100 amid acids. The synonymous substitution rate (*Ks*) of reciprocal best hit gene pairs was calculated based on the YN model in *KaKs_Calculator* v2.0⁷². Synteny analysis was performed on *Taxus* protein-coding genes using MCSanX⁷³ to identify whole genome duplication events with default parameters from the top ten self-blastp hits. The synonymous substitution rate (*Ks*) and 4DTv were calculated for *Taxus* syntenic block gene pairs.

Genome mining for CYP450s and gene clusters

For the identification and classification of CYP450 genes, hmmsearch was used to identify CYP450 genes in the *Taxus* genome with PF00067 from the Pfam database⁷⁴. Classification of the 649 CYP450 genes was executed by alignment with the CYP450 database⁷⁵ using standard sequence similarity cutoffs, with definite standards of 97%, 55% and 40% for allelic, subfamily and family variants, respectively. According to the standardized CYP450 nomenclature⁷⁶, CYP450s were divided into A-type and non-A-type CYP450s, and phylogenetic analysis of CYP450 genes was performed for A-type and non-A-type CYP450s. Neighbor-joining

phylogenetic trees were constructed using the MEGA7 package with homologous amino acid sequences⁷⁷.

For genome mining for gene clusters involved in plant specialized metabolism, PlantiSMASH³¹ was used to search for potential gene clusters using the default parameters and the GFF (General Feature Format) annotation files of the software. Gene groups were identified by in silico analysis based on the following criteria: 1) the distance between two adjacent CYP450 genes in one group should be less than 5.26 Mb; and 2) one group should contain at least 7 CYP450 genes.

RNA-seq data analysis for candidate genes in the paclitaxel biosynthesis pathway

All tissues, including female strobilus (mfs), leaf (mfl), bark of stem (mfb), and root (mfr) and male strobili (mtf), leaf (mtl), bark of stem (mtb), and root (mtr), were mapped to the *Taxus* genome, and the FPKM value was calculated using HISAT2 and StringTie⁷⁶. Expression data from female bark, female roots and female leaves were used to identify the genes associated with paclitaxel biosynthesis. First, we selected genes that were highly expressed in roots or bark compared to leaves. Second, the genes were further confirmed in two *Taxus* half-sib cell lines (HC and LC) with distinct accumulation patterns of paclitaxel. The genes should be significantly highly expressed in HC. The differentially expressed genes were filtered using edgeR⁷⁷ with $\log_{2}FC > 1$ and $FDR < 0.05$. Finally, we obtained 1,638 genes that met the thresholds above. Gene-to-gene networks were constructed using the expression matrix from MeJA-induced cell line (0, 2, 4, 8, 24 h) RNA-seq data. Pearson correlation analysis was performed with the known functional genes as the target genes. Hypothesis development for the Pearson correlation was performed, and pairs with a P-value < 0.05 remained.

Functional characterization of TS genes

The ORFs (open reading frames) of *TS1* (*ctg6088_gene.1*) and *TS2* (*ctg5306_gene.4*) were cloned by reverse transcription PCR (RT-PCR) from the HC cell line. Plant-Ploc (<http://www.csbio.sjtu.edu.cn/bioinf/plant-multi/>), ChloroP (<http://www.cbs.dtu.dk/services/ChloroP/>), and TargetP (<http://www.cbs.dtu.dk/services/TargetP/>) were used for the prediction of the plastidial target sequence. The 60-residue N-terminally truncated *TS1* and *TS2* genes were inserted into the *Escherichia coli* (*E. coli*) expression vector pET28b to form the constructs pET28b::*TS1* and pET28b::*TS2*, respectively. All expression plasmids were constructed using the Hieff Clone™ One Step Cloning Kit (YEASEN, China) and the primers used in this work are given in Supplementary Table 28. For the *in vitro* enzyme assay, the enzyme assays were performed in a final volume of 500 μ L buffer (25 mM HEPES, pH 8.5, 10% glycerol, 5 mM DTT, 5 mM sodium ascorbate, 5 mM sodium metabisulfite and 1 mM $MgCl_2$) containing 100 μ g of purified protein and 100 μ M GGPP (Sigma-Aldrich). The reaction mixture was overlaid with 500 μ L of pentane (Macklin, GC-MS grade) and incubated overnight at 32°C. In addition, the mixture was vortexed, and the pentane overlay was subsequently removed by centrifugation at 5,000 rpm for 10 min and concentrated by N_2 gas before gas chromatography mass spectrometry (GC/MS) analysis. Inactivated TSs-His6 was used as the control. Taxa-4(5),11(12)-diene (**1**) and taxa-4(20),11(12)-diene (**2**) preparations were performed according to a previous study with the taxadiene-producing *E. coli* strain T2 (harboring pMH1, pFZ81, and pXC02)³⁷. The organic solutions containing crude compounds **1** and **2** were concentrated on ice under N_2 gas and re-dissolved in dimethyl sulfoxide (DMSO) for the purification of compounds **1** and **2** by thin layer chromatography (TLC). The purity and concentration were determined by GC-MS (Supplemental Fig. 25).

For the determination of kinetic parameters, standard enzyme assays were carried out in a total volume of 100 μ L containing buffer (25 mM HEPES, pH 8.5, 10% glycerol, 5 mM DTT, 5 mM sodium ascorbate, 5 mM sodium metabisulfite and 1 mM MgCl_2), 34 μ g (TS1) or 18 μ g (TS2) of recombinant proteins and seven different concentrations of GGPP (0.2, 0.5, 1, 2.5, 5, 10, 25 and 50 μ M), which were spiked with [$^1\text{-}^3\text{H}$]-GGPP (American Radiolabeled Chemicals, Inc, 30 Ci mM^{-1}). Hot [$^1\text{-}^3\text{H}$]-GGPP was diluted 400 times using cool GGPP (Sigma, 1 mg mL^{-1}). The reaction mixtures were incubated at 32°C for 30 min and then quenched for 10 min using 100 μ L of stop solution (containing 1 M EDTA and 4 M NaOH). The reaction mixture was extracted with 800 μ L of n-hexane (vortexed for 10 s and 12,000 rpm for 2 min), and 400 μ L of the n-hexane layer was subsequently removed and mixed with 2 mL of the liquid scintillation cocktail. The total radioactivity of the reaction products was measured using a liquid scintillation counter (Tri-Carb 2910TR, Perkin Elmer). The kinetic constant was calculated by a nonlinear regression fit to the Michaelis-Menten equation using OriginPro® 8.6 (OriginLab Co., Northampton)⁷⁸.

E. coli TS2 was constructed by replacing *TbTS* with *TS2* based on the previous taxadiene-producing *E. coli* *TbTS* (harboring pMH1, pFZ81, and pXC02, and coexpressing nine genes *AtoB*, *ERG13*, *tHMG1*, *ERG12*, *ERG8*, *MVD1*, *IdI*, *GGPPS* and *TbTS* in *E. coli*)³⁷ (Supplementary Fig. 20a). The *E. coli* strains T2 and TS2 were cultivated in 50-mL flasks containing 30 mL of LB medium at 37°C with 100 mg L^{-1} ampicillin (AMP), 50 mg L^{-1} kanamycin (KAN), and 34 mg L^{-1} chloramphenicol (CM). When the OD_{600} reached approximately 0.1, 1 mM isopropyl β -D-1-thiogalactopyranoside (IPTG) was added to the cultures along with 3 mL of dodecane, and then, the bacteria were cultivated at 28°C. Experiments were repeated four times. For cell concentration (OD_{600}) and taxadiene

measurement, 100-μL cultures and 30-μL organic layers were collected at set intervals (at 8, 13, 22, 37, 48, 60 and 72 hours). The produced taxadiene was detected with GC-MS and quantified with the nonyl acetate standard (Aladdin).

References

- 1 Christenhusz, M. *et al.* A new classification and linear sequence of extant gymnosperms. *Nov. Magnolia Press Phytotaxa* **19**, 55-70, doi:10.11646/phytotaxa.19.1.3 (2010).
- 2 Hao, D. C., Xiao, P. G., Huang, B., Ge, G. B. & Yang, L. 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, doi:10.1007/s00606-008-0069-0 (2008).
- 3 Wani, M. C., Taylor, H. L., Wall, M. E., Coggon, P. & McPhail, A. T. Plant antitumor agents. VI. Isolation and structure of taxol, a novel antileukemic and antitumor agent from *Taxus brevifolia*. *J. Am. Chem. Soc.* **93**, 2325-2327, doi:10.1021/ja00738a045 (1971).
- 4 Sabzehzari, M., Zeinali, M. & Naghavi, M. R. Alternative sources and metabolic engineering of Taxol: Advances and future perspectives. *Biotechnol. Adv.* **43**, 107569, doi:10.1016/j.biotechadv.2020.107569 (2020).
- 5 Nicolaou, K. C. *et al.* Total synthesis of taxol. *Nature* **367**, 630-634, doi:10.1038/367630a0 (1994).
- 6 Baloglu, E. & Kingston, D. G. I. A new semisynthesis of paclitaxel from baccatin III. *J. Nat. Prod.* **62**, 1068-1071, doi:10.1021/np990040k (1999).

- 7 Fett-Neto, A. G., DiCosmo, F., Reynolds, W. F. & Sakata, K. Cell culture of *Taxus* as a source of the antineoplastic drug taxol and related taxanes. *Nat. Biotechnol.* **10**, 1572-1575, doi:10.1038/nbt1292-1572 (1992).
- 8 Kumar, P. *et al.* Hyper-production of taxol from *Aspergillus fumigatus*, an endophytic fungus isolated from *Taxus* sp. of the Northern Himalayan region. *Biotechnol Rep (Amst)* **24**, e00395-e00395, doi:10.1016/j.btre.2019.e00395 (2019).
- 9 Ajikumar, P. K. *et al.* Isoprenoid pathway optimization for Taxol precursor overproduction in *Escherichia coli*. *Science* **330**, 70-74, doi:10.1126/science.1191652 (2010).
- 10 Kuang, X., Sun, S., Wei, J., Li, Y. & Sun, C. Iso-Seq analysis of the *Taxus cuspidata* transcriptome reveals the complexity of taxol biosynthesis. *BMC Plant Biol.* **19**, 210-226, doi:10.1186/s12870-019-1809-8 (2019).
- 11 Croteau, R., Ketchum, R. E. B., Long, R. M., Kaspera, R. & Wildung, M. R. Taxol biosynthesis and molecular genetics. *Phytochem. Rev.* **5**, 75-97, doi:10.1007/s11101-005-3748-2 (2006).
- 12 Wildung, M. R. & Croteau, R. A cDNA clone for taxadiene synthase, the diterpene cyclase that catalyzes the committed step of taxol biosynthesis. *J. Biol. Chem.* **271**, 9201-9204, doi:10.1074/jbc.271.16.9201 (1996).
- 13 Howat, S. *et al.* Paclitaxel: biosynthesis, production and future prospects. *N. Biotechnol.* **31**, 242-245, doi:10.1016/j.nbt.2014.02.010 (2014).
- 14 Sanchez-Muñoz, R. *et al.* A novel hydroxylation step in the taxane biosynthetic pathway: A new approach to paclitaxel production by synthetic biology. *Front. Bioeng. Biotech.* **8**, doi:10.3389/fbioe.2020.00410 (2020).

- 15 Walker, K. & Croteau, R. Taxol biosynthesis: Molecular cloning of a benzoyl-CoA:taxane 2 α -O-benzoyltransferase cDNA from *Taxus* and functional expression in *Escherichia coli*. *P. Natl. Acad. Sci. USA* **97**, 13591-13596, doi:10.1073/pnas.250491997 (2000).
- 16 Fett-Neto, A. G., Melanson, S. J., Sakata, K. & DiCosmo, F. Improved growth and taxol yield in developing calli of *taxus cuspidata* by medium composition modification. *Nat. Biotechnol.* **11**, 731-734, doi:10.1038/nbt0693-731 (1993).
- 17 Wasternack, C. Action of jasmonates in plant stress responses and development--applied aspects. *Biotechnol. Adv.* **32**, 31-39, doi:10.1016/j.biotechadv.2013.09.009 (2014).
- 18 Cusido, R. M. *et al.* A rational approach to improving the biotechnological production of taxanes in plant cell cultures of *Taxus spp.* *Biotechnol. Adv.* **32**, 1157-1167, doi:10.1016/j.biotechadv.2014.03.002 (2014).
- 19 Zhang, M. *et al.* Transcriptome-wide identification and screening of WRKY factors involved in the regulation of taxol biosynthesis in *Taxus chinensis*. *Sci. Rep.* **8**, 5197-5209, doi:10.1038/s41598-018-23558-1 (2018).
- 20 Lenka, S. K. *et al.* Jasmonate-responsive expression of paclitaxel biosynthesis genes in *Taxus cuspidata* cultured cells is negatively regulated by the bHLH transcription factors TcJAMYC1, TcJAMYC2, and TcJAMYC4. *Front. Plant Sci.* **6**, 115-129, doi:10.3389/fpls.2015.00115 (2015).
- 21 Yu, C. *et al.* Tissue-specific study across the stem of *Taxus media* identifies a phloem-specific TmMYB3 involved in the transcriptional regulation of paclitaxel biosynthesis. *Plant J.* **10**, tjp14710, doi:10.1111/tpj.14710 (2020).

- 22 Zonneveld, B. J. M. Conifer genome sizes of 172 species, covering 64 of 67 genera, range from 8 to 72 picogram. *Nord. J. Bot.* **30**, 490-502, doi:10.1111/j.1756-1051.2012.01516.x (2012).
- 23 Li, Y. *et al.* A protocol of homozygous haploid callus induction from endosperm of *Taxus chinensis* Rehd. var. *mairei*. *SpringerPlus* **5**, 659-668, doi:10.1186/s40064-016-2320-4 (2016).
- 24 Jiao, Y. *et al.* Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**, 97-100, doi:10.1038/nature09916 (2011).
- 25 Nystedt, B. *et al.* The Norway spruce genome sequence and conifer genome evolution. *Nature* **497**, 579-584, doi:10.1038/nature12211 (2013).
- 26 Zhang, Q. J. & Gao, L. Z. Rapid and recent evolution of LTR retrotransposons drives rice genome evolution during the speciation of AA-genome *Oryza* species. *G3* **7**, 1875-1885, doi:10.1534/g3.116.037572 (2017).
- 27 Guerra-Bubb, J., Croteau, R. & Williams, R. M. The early stages of taxol biosynthesis: An interim report on the synthesis and identification of early pathway metabolites. *Nat. Prod. Rep.* **29**, 683-696 (2012).
- 28 Gesell, A. *et al.* The gymnosperm cytochrome P450 CYP750B1 catalyzes stereospecific monoterpene hydroxylation of (+)-Sabinene in thujone biosynthesis in western redcedar. *Plant Physiol.* **168**, 94-106, doi: 10.1104/pp.15.00315 (2015).
- 29 Nelson, D. & Werck-Reichhart, D. A P450-centric view of plant evolution. *Plant J.* **66**, 194-211, doi:10.1111/j.1365-3113.2011.04529.x (2011).

- 30 Wasternack, C. & Strnad, M. Jasmonates are signals in the biosynthesis of secondary metabolites — pathways, transcription factors and applied aspects — a brief review. *N. Biotechnol.* **48**, 1-11, doi:10.1016/j.nbt.2017.09.007 (2019).
- 31 Kautsar, S., Suarez, H., Blin, K., Osbourn, A. & Medema, M. PlantSMASH: Automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Res.* **45**, doi:10.1093/nar/gkx305 (2017).
- 32 Shang, Y. *et al.* Biosynthesis, regulation, and domestication of bitterness in cucumber. *Science* **346**, 1084-1088, doi:10.1126/science.1259215 (2014).
- 33 Ben, F. & Anne, E. O. Metabolic diversification—-independent assembly of operon-like gene clusters in different plants. *Science* **320**, 543, doi:10.1093/pcp/pcq063 (2008).
- 34 Bertoli, D. *et al.* The genome sequence of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nat. Genet.* **48**, doi:10.1038/ng.3517 (2016).
- 35 Guan, R. *et al.* Draft genome of the living fossil *Ginkgo biloba*. *Gigascience* **5**, 49, doi:10.1186/s13742-016-0154-1 (2016).
- 36 De La Torre, A., Li, Z., Van de Peer, Y. & Ingvarsson, P. Contrasting rates of molecular evolution and patterns of selection among gymnosperms and flowering plants. *Molecular Biol. Evol.* **34**, doi:10.1093/molbev/msx069 (2017).
- 37 Bian, G. *et al.* Production of taxadiene by engineering of mevalonate pathway in *Escherichia coli* and endophytic fungus *Alternaria alternata* TPF6. *Biotechnol. J.* **12**, doi:10.1002/biot.201600697 (2017).
- 38 Li, Y. *et al.* Induction of half-sib embryonic callus and production of taxiod compounds from *Taxus chinensis* var. *mairei*. *Int. J. Agric. Biol.* **21**, 719-725, doi:10.17957/IJAB/15.0949 (2019).

- 39 Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat. Methods* **17**, 155-158, doi:10.1038/s41592-019-0669-3 (2020).
- 40 Wang, J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963, doi:10.1371/journal.pone.0112963 (2014).
- 41 Servant, N. *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259, doi:10.1186/s13059-015-0831-x (2015).
- 42 Durand, N. *et al.* Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95-98, doi:10.1016/j.cels.2016.07.002 (2016).
- 43 Dudchenko, O. *et al.* De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92-95, doi:10.1126/science.aal3327 (2017).
- 44 Burton, J. N. *et al.* Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119-1125, doi:10.1038/nbt.2727 (2013).
- 45 Xu, M. *et al.* TGS-GapCloser: A fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *GigaScience* **9**, doi:10.1093/gigascience/giaa094 (2020).
- 46 Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC bioinform.* **19**, 460, doi:10.1186/s12859-018-2485-7 (2018).
- 47 Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357-360, doi:10.1038/nmeth.3317 (2015).
- 48 Bao, W., Kojima, K. K. & Kohany, O. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11, doi:10.1186/s13100-015-0041-9 (2015).

- 49 Ouyang, S. & Buell, C. R. The TIGR Plant Repeat Databases: a collective resource for
the identification of repetitive sequences in plants. *Nucleic Acids Res.* **32**, D360-363,
doi:10.1093/nar/gkh099 (2004).
- 50 Ou, S. & Jiang, N. LTR_retriever: A highly accurate and sensitive program for
identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410-1422,
doi:10.1104/pp.17.01310 (2018).
- 51 Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software
for de novo detection of LTR retrotransposons. *BMC bioinform.* **9**, 18, doi:10.1186/1471-
2105-9-18 (2008).
- 52 Korf, I. Gene finding in novel genomes. *BMC Bioinform.* **5**, 59, doi:10.1186/1471-2105-
5-59 (2004).
- 53 Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron
submodel. *Bioinformatics* **19**, ii215-ii225, doi:10.1093/bioinformatics/btg1080 (2003).
- 54 The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering
plant *Arabidopsis thaliana*. *Nature* **408**, 796-815, doi:10.1038/35048692 (2000).
- 55 Goff, S. A. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. *ssp.* *japonica*).
Science **296**, 92-100, doi:10.1126/science.1068275 (2002).
- 56 Wan, T. *et al.* A genome for gnetophytes and early evolution of seed plants. *Nat. Plants* **4**,
82-89, doi:10.1038/s41477-017-0097-2 (2018).
- 57 Banks, J. A. *et al.* The Selaginella genome identifies genetic changes associated with the
evolution of vascular plants. *Science* **332**, 960-963, doi:10.1126/science.1203810 (2011).
- 58 Zimin, A. *et al.* Sequencing and assembly of the 22-Gb loblolly pine genome. *Genetics*
196, 875-890, doi:10.1534/genetics.113.159715 (2014).

- 59 Amborella Genome Project. The Amborella genome and the evolution of flowering
plants. *Science* **342**, 1241089, doi:10.1126/science.1241089 (2013).
- 60 Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988-
995, doi:10.1101/gr.1865504 (2004).
- 61 Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of
insertions, deletions and gene fusions. *Genome Biol.* **14**, R36, doi:10.1186/gb-2013-14-4-
r36 (2013).
- 62 Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq
experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562-578,
doi:10.1038/nprot.2012.016 (2012).
- 63 Haas, B. J. Improving the Arabidopsis genome annotation using maximal transcript
alignment assemblies. *Nucleic Acids Res.* **31**, 5654-5666, doi:10.1093/nar/gkg770 (2003).
- 64 Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler
and the Program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).
- 65 Blake, J. A., Chan, J., Kishore, R., Sternberg, P. W. & Li, Y. Gene Ontology Consortium:
going forward. *Nucleic Acids Res.* **43**, 1049-1056, doi:10.1093/nar/gku1179 (2015).
- 66 Mitchell, A. L. *et al.* InterPro in 2019: improving coverage, classification and access to
protein sequence annotations. *Nucleic Acids Res.* **47**, 351-360, doi:10.1093/nar/gky1100
(2019).
- 67 Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence data bank and its new
supplement TrEMBL. *Nucleic Acids Res.* **24**, 21-25, doi:10.1093/nar/24.1.21 (1996).
- 68 Stoesser, G., Sterk, P., Tuli, M. A., Stoehr, P. J. & Cameron, G. N. The EMBL nucleotide
sequence database. *Nucleic Acids Res.* **25**, 7-13, doi:10.1093/nar/25.1.7 (1997).

- 69 Lamesch, P. *et al.* The Arabidopsis Information Resource (TAIR): improved gene
annotation and new tools. *Nucleic Acids Res.* **40**, 1202-1210, doi:10.1093/nar/gkr1090
(2012).
- 70 Bailey, J. A., Church, D. M., Ventura, M., Rocchi, M. & Eichler, E. E. Analysis of
segmental duplications and genome assembly in the mouse. *Genome Res.* **14**, 789-801,
doi:10.1101/gr.2238404 (2004).
- 71 Putnam, N. H. *et al.* The amphioxus genome and the evolution of the chordate karyotype.
Nature **453**, 1064-1071, doi:10.1038/nature06967 (2008).
- 72 Wang, D. P., Wan, H. L., Zhang, S. & Yu, J. γ -MYN: a new algorithm for estimating Ka
and Ks with consideration of variable substitution rates. *Biol. Direct* **4**, 20,
doi:10.1186/1745-6150-4-20 (2009).
- 73 Wang, Y. *et al.* MCSScanX: a toolkit for detection and evolutionary analysis of gene
synteny and collinearity. *Nucleic Acids Res.* **40**, 49-63, doi:10.1093/nar/gkr1293 (2012).
- 74 Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput Biol* **7**, e1002195,
doi:10.1371/journal.pcbi.1002195 (2011).
- 75 Nelson, D. R. The cytochrome p450 homepage. *Hum. Genomics* **4**, 59-65,
doi:10.1186/1479-7364-4-1-59 (2009).
- 76 Durst, F. & Nelson, D. R. Diversity and evolution of plant P450 and P450-reductases.
Drug Metabol Drug Interact **12**, 189-206, doi:10.1515/dmdi.1995.12.3-4.189 (1995).
- 77 Kumar, S., Stecher, G. & Tamura, K. MEGA7: molecular evolutionary genetics analysis
version 0 for bigger datasets. *Mol. Biol. Evol.* **33**, 1870-1874,
doi:10.1093/molbev/msw054 (2016).
- 78 Grant, F. Origin Pro 8.6. *Scientific Computing World* (2011).

Acknowledgments

We particularly thank Tiangang Liu, from Wuhan University, for sharing plasmids that were used for the preparation of taxadiene in *E. coli*. We thank Yansheng Zhang (Shanghai University), and Shifeng Cheng (Agricultural Genomics Institute at Shenzhen) for helpful discussion. We are grateful for valuable help with TS evolution from Tao Feng (Wuhan Botanical Garden, Chinese Academy of Sciences). **Funding:** This work was supported by the National Key R&D Program of China (2018YFA0903200, 2018YFA0901800), Research Funds for Central Nonprofit Scientific Institution (Y2020XK23), the Elite Young Scientists Program of CAAS, the Agricultural Science and Technology Innovation Program, National Science and Technology Basic Special Project (2017FY100100), the Scientific Research Fund of Hunan Provincial Education Department (2016XYX001), Double First-class Construction Project of Hunan Agricultural University (SYL201802026), Fund of the Education Department of Hunan Province (18B124), and the National Natural Science Foundation of China (32000236).

Author contributions

The ideas for the paper were conceived by J.Y., S.H., and X.X.; Q.L., G.B., Q.Z., X.W., T.S., Z.Z., and Y.W. performed bioinformatic analysis; Y.L., X.X., C.L., D.R., L.G., and P.L. prepared the *Taxus* sequencing samples and the *Taxus* cell lines; J.G., H.L., D.R., Y.S. and R.D. undertook enzymatic analysis and mass spectrometry; J.Y., H.S., J.G., Q.L., Y.L. and X.X. interpreted the data and wrote the paper.

804

805 **Competing interests**

806 The authors declare no competing interests.

807

808 **Data and materials availability**

809 The raw data used in this study have been deposited in the Genome Sequence Archive at the BIG

810 Data Center, Beijing Institute of Genomics (BIG), Chinese Academy of Sciences and are

811 accessible at <http://bigd.big.ac.cn/gsa> under bioproject PRJCA003841.

Fig. 1. Genomic features of *T. chinensis* var. *mairei*.

- (a) Genomic landscape of the twelve assembled pseudochromosomes. Track a represents the length of the pseudochromosomes (Mb); b, c, and d represents repeat element density, GC content, and distribution of gene density, respectively; e, f and g show the distribution of Ty3/Gypsy, Ty1/Copia, and unknown LTRs, respectively. These metrics are calculated in 5 Mb windows.
- (b) Whole genome duplication (WGD) analysis based on the substitution rate distribution of paralogs. Upper panel: Histogram of the *Ks* distribution from *Taxus* paralogs based on an all-to-all blast to total genes. Lower panel: *Ks* distribution of paralogs based on syntenic analysis. *Ks* values were calculated using the YN model in *KaKs_calculator*.
- (c) Expansions and diverse sets of LTR elements in the *Taxus* genome. The histogram shows distributions of insertion times calculated for LTRs in *Taxus* and rice, using mutation rates (per base year) of 7.3×10^{-10} for *Taxus* and 1.8×10^{-8} for rice. The LTR-RT insertions of *T. chinensis* var. *mairei* and *Z. mays* are shown as columns in different colors.
- (d) Heuristic maximum like hood trees of Ty3/Gypsy (shown as Gypsy) and Ty1/Copia (shown as Copia) from 6 plant species. The two trees were constructed from amino acid sequence similarities within the reverse transcriptase domains of Gypsy and Copia from 6 plant species. Gypsy elements are divided into eight families (I-VIII), and Copia contains five families (I-V). The representative plants are shown as colored lines.
- (e) Venn diagram for orthologous protein-coding gene clusters in cryptogam (Cry), angiosperm (Ang), gymnosperm (Gym), and *T. chinensis* var. *mairei* (Tax). The cryptogams include *M. polymorpha*, *P. patens* subsp. *patens*, and *S. moellendorffii*. The angiosperms include *A. trichopoda*, *V. vinifera*, *A. thaliana*, *S. miltiorrhiza*, and *O. sativa*. The gymnosperms include *P. abies* and *G. biloba*. The number in each sector of the diagram represents the total number of genes across the four comparisons.
- (f) Evolution analysis of gene families in *Taxus* and selected plants. The red numbers on the branches of the phylogenetic tree indicate the number of expanded gene families, and the blue numbers refer to the number of constricted gene families. The supposed most recent common ancestor (MRCA) contains 26, 974 gene families. G, L, E, and C in the table at right represent the number of gains, losses, expansions and constrictions in the gene families among 11 plant species.

Fig. 2 Evolution and genomic architecture of *Taxus* CYP450s.

(a) Phylogenomic analysis of the non-A-type CYP450s in the representative plant species. Abbreviations: A, Angiosperms; G, Gymnosperms; P, Pteridophytes; B, Bryophytes. The color of each block is based on the number of genes in each family, and 0, 1, 2, and 3 indicate that the ranges from 0, 1-10, 10-50, and 50-100 genes, respectively.

(b) Phylogenetic analysis of the CYP725 subfamily in *T. chinensis* var. *mairei* (Taxus), *G. biloba* (Ginkgo), *P. abies* (Picea), and *C. revolute* (Cycas). The dotted outline shows the gene spheres of the CYP725A and CYP725B subfamilies. The dusty blue dots on the end of

the phylogenetic branches represent the known paclitaxel pathway CYP725A genes and their homologs. The neighbor-joining tree was constructed by Interactive Tree Of Life (iTOL) software. The evolutionary distances were analyzed by the *p*-distance method, and the branch lengths were scaled by the bar.

- (c) Distribution of CYP450 genes on the twelve pseudochromosomes in *Taxus*. Each short line on the pseudochromosomes represents a CYP450 gene. CYP725As, CYP725Bs, and the other CYP450s are marked by red, orange and gray lines, respectively. The known CYP450s in the paclitaxel biosynthesis pathway (known CYP) are shown in blue. The CYP450 groups (≥ 7 CYP450 genes and ≤ 5.26 Mb of gene spacing between two adjacent CYP450s) are labeled outside of the corresponding position on the pseudochromosomes.
- (d) Histogram plot of the number of CYP450 genes on each pseudochromosome. The CYP725 genes (shown in red and orange) were mainly distributed on pseudochromosome 9, while the other CYPs (shown in grey) were distributed randomly on 12 pseudochromosomes. The Y-axis represents the number of CYP450 genes.
- (e) Group-based gene expression profiles in response to methyl jasmonate (MeJA) treatment. RNA sequencing analysis was performed with the low paclitaxel-yielding cell line (LC) treated with 100 μ M MeJA for 4 h. The expression of the gene group was calculated by the sum of the expression levels of each CYP450, and each upregulated and downregulated CYP450 was calculated as 1 and -1, respectively, based on their RPKM (reads per kilobase per million reads) values.
- (f) Map of CYP725As located in groups 9.1 and 9.2. The ranges of the gene groups on pseudochromosome 9 is marked in pink. CYP725As and the other genes are marked by red and gray vertical lines, respectively. The known CYP450s in the paclitaxel biosynthesis pathway (known CYP) are shown in blue. The arrows show gene orientations.

Fig. 3. Functional identification of the paclitaxel biosynthesis gene cluster.

(a) Genomic architecture and expression pattern of the taxadiene cluster. The arrows indicate the relative position and direction of the genes in the cluster. Here, 55,305,455 and 55,566,094 indicate the starting and ending positions of the cluster on pseudochromosome 9. The two

unknown CYP725A genes are represented by their gene starting position 55326109 and 55305455, respectively. TS1 and T5 α H3 are located at 72,105,619-72,109,598 and 49,866,845-49,868,629 bp on chromosome 9, respectively. The relative expression levels of taxadiene cluster genes in *Taxus* are based on their RPKM values. The expression level of genes with high sequence similarity was distinguished based on sequencing read counts of the exons that include different bases, and adjusting the alignment threshold to no mismatch. RNA-Seq datasets are from roots, leaves, and bark of male plants (shown in green); two *T. chinensis* var. *mairei* half-sib cell lines, a high paclitaxel-yielding cell line (HC) and a low paclitaxel-yielding line (LC), respectively (shown in yellow); and MeJA-treated LC (+MJ) and MeJA-untreated LC (-MJ) (shown in orange). Data are means $\pm$ S.D. (n = 3).

- (b) Analysis of TS activity *in vitro*. The purified recombinant TS1-His and TS2-His were incubated with the substrate GGPP overnight at 32°C. The reaction products were analyzed by GC/MS. TS catalyzes GGPP to produce a major product (**1**, taxa-4(5),11(12)-diene) and a minor product (**2**, taxa-4(20),11(12)-diene), while boiled TSs have no TS activity. *m/z* 122 is a characteristic ion of taxadienes. The taxadiene confirmed by NMR analysis was used as a reference standard (Standard).
- (c) Analysis of the activity of T5 α H and two novel CYP725As *in vitro*. The *in vitro* enzyme assay was carried out with the purified taxadiene substrate and yeast microsomes, each including one of the six CYPs (T5 α H1, T5 α H2, T5 α H3, TbT5 α H, 55326109 or 55305455) and cytochrome P450 reductase (CPR). T5 α H1/2/3 can produce three oxygenated taxadiene products (**3**, 5(12)-oxa-3(11)-cyclotaxane (OCT); **4**, 5(11)-oxa-3(11)-cyclotaxane (iso-OCT), and **5**, taxa-4(20),11(12)-dien-5 α -ol), whereas no catalytic compounds were observed for 55326109, 55305455 and CPR. *T. brevifolia* taxadiene 5- α -hydroxylase (TbT5 α H), shown to have taxadiene 5- α -hydroxylase activity, was used as a positive control.
- (d) Kinetic evaluation of GGPP oxidation catalyzed by TS1 (blue circle) and TS2 (red rectangle). $K_m = 5.5 \pm 1.6 \mu\text{M}$ (TS1), $K_m = 8.6 \pm 1.5 \mu\text{M}$ (TS2), $k_{\text{cat}} = 0.0029 \text{ s}^{-1}$ (TS1), and $k_{\text{cat}} = 0.0090 \text{ s}^{-1}$ (TS2).
- (e) Quantitative real-time PCR analysis of the transcription levels of TS1 and TS2 in the *Taxus* cell line LC treated with 100 μM MeJA for the indicated times. The relative gene expression levels are represented as the average fold change ($2^{-\Delta\Delta C_t}$). The *Taxus* actin 1 gene (7G702435613) was used as an internal reference. Data are means $\pm$ S.D. (n = 3).

(f) Biosynthesis pathway of paclitaxel in *T. chinensis* var. *mairei*. The solid arrows indicate the identified steps in the paclitaxel pathway, whereas the dotted arrows show the hypothetical steps. TS, taxadiene synthase; T5 α H, taxadiene 5- α -hydroxylase; T13 α H, taxane 13- α -hydroxylase; TAT, taxadien-5- α -ol O-acetyltransferase; T10 β H, taxane 10- β -hydroxylase; T14 β H, taxoid 14- β -hydroxylase; T2 α H, taxoid 2- α -hydroxylase; T7 β H, taxoid 7- β -hydroxylase; TBT, 2- α -hydroxytaxane 2-O-benzoyltransferase; DBAT, 10-deacetylbaccatinIII 10-O-acetyltransferase; BAPT, baccatin III amino phenylpropanoyl-13-O-transferase; and DBTNBT, 3'-N-debenzoyl-2'-deoxytaxol N-benzoyl transferase.