

Archipelago architecture and evolution of paclitaxel biosynthesis revealed by the *Pseudotaxus chienii* haplotype-resolved genome

Received: 20 December 2025

Accepted: 14 August 2026

Published online: 26 August 2026



Huan Wang^{1,2,7}, Haidong Chen^{1,7}, Yaping Sun^{1,7}, Meifeng Su¹, Yi Zhang^{1,3,4}, Wangyang Xie¹, Ran Du¹, Yugeng Liu¹, Yuchen Xu¹, Shengchao Liu¹, Bin Jiang^{1,5}, Shaoxing Bai¹, Yanchun Peng¹, Lianming Gao⁶, Jiaming Li¹✉ & Jianbin Yan¹✉

Pseudotaxus chienii, the only species in the genus *Pseudotaxus*, is closely related to *Taxus* genus. Whether *P. chienii* can produce paclitaxel has long remained controversial. Here, we present a chromosome-level, haplotype-resolved genome of *P. chienii*, with each haplotype (~15.4 Gb) exhibiting high continuity and completeness. We identify a structurally conserved yet physically dispersed genomic region, termed the Taxane Biosynthetic Archipelago (TBA), containing three biosynthetic gene clusters and collinear core genes essential for taxane biosynthesis. Integrative transcriptomic, metabolite, and heterologous enzymatic analyses reveal that while *P. chienii* retains a largely conserved genomic architecture associated with paclitaxel biosynthesis, its terminal pathway shows severe functional degradation and may have largely lost the functional capacity for paclitaxel production. Concurrently, it exhibits significant accumulation of taxinine J. This study provides an evolutionary framework for interpreting the long-standing debate on paclitaxel production beyond *Taxus*, and understanding the evolutionary decay of complex specialized pathways.

Pseudotaxus chienii, a rare endemic understory conifer, is characterized by seeds enclosed in a white fleshy aril, distinct from the typically red arils of *Taxus* (Supplementary Fig. 1)¹. It occupies moist rocky habitats in subtropical montane forests at 600–1500 m elevation, with fragmented distributions confined to isolated populations in south-eastern China. Due to habitat fragmentation, it is on the International Union for Conservation of Nature (IUCN) vulnerable and endangered

list^{2,3}. As the sole species in the monotypic genus *Pseudotaxus* (*Taxaceae*)⁴, it serves as an important model for understanding the evolutionary diversification of *Taxaceae*.

In particular, the phylogenetic and morphological similarities between *P. chienii* and *Taxus* make it a suitable option for comparative studies on taxoid metabolism. *Taxus* species yield paclitaxel (Taxol), a vital diterpenoid for cancer chemotherapy and among the most

¹Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Key Laboratory of Synthetic Biology, Ministry of Agriculture and Rural Affairs, Advanced Synthetic Biology Institute at Hohhot, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China.

²Guangdong Province Key Laboratory of Microbial Signals and Disease Control, Integrative Microbiology Research Centre, and College of Plant Protection, South China Agricultural University, Guangzhou, China. ³School of Life Sciences, Henan University, Kaifeng, China. ⁴Shenzhen Research Institute of Henan University, Shenzhen, China. ⁵MOE Key Laboratory of Bioinformatics, Tsinghua-Peking Center for Life Sciences, School of Life Sciences, Tsinghua University, Beijing, China. ⁶Center for Interdisciplinary Biodiversity Research & College of Forestry, Shandong Agricultural University, Tai'an, China. ⁷These authors contributed equally: Huan Wang, Haidong Chen, Yaping Sun.

✉ e-mail: lijiaming@caas.cn; jianbinlab@caas.cn

studied plant specialized metabolites^{5,6}. The paclitaxel biosynthesis entails clustered, co-regulated enzymes, including cytochrome P450s, acyltransferases, and hydroxylases. Genomic analyses in *Taxus mairei*⁷, *T. wallichiana*⁸, and *T. yunnanensis*⁹ have uncovered novel taxadiene synthases, conserved biosynthetic gene clusters, and provided evolutionary insights. Despite these findings, gymnosperms generally possess exceptionally large and highly repetitive genomes, posing substantial challenges for resolving their structural organization at high resolution. Consequently, it remains poorly understood how their complex, multi-step specialized metabolic pathways in gymnosperms preserve genomic architectural integrity while simultaneously undergoing evolutionary diversification. Thus, *P. chienii* represents an ideal model to dissect the origin, genomic architectural conservation, and evolutionary divergence of the paclitaxel biosynthetic pathway across *Taxaceae*.

Although significant progress has been made in elucidating its biosynthetic pathway in *Taxus* species^{7–13}, research on paclitaxel biosynthesis in *P. chienii* is still limited and presents conflicting results. Previous studies report conflicting paclitaxel levels in *P. chienii*, with none detected by Heinig et al.¹⁴ but substantial accumulation noted by Yu et al.¹⁵, highlighting the need for comprehensive genomic investigation to resolve these discrepancies. A recent study by Wang et al.¹⁶ further intensified this controversy by proposing that *P. chienii* possesses only an incomplete taxane biosynthetic pathway, which terminates before the synthesis of 10-deacetylbaicatin III. However, conclusions derived from conventional consensus genome assemblies may be inherently limited in highly heterozygous species. Accurate characterization of complex specialized metabolic pathways requires haplotype-resolved genome assemblies, which enable comprehensive characterization of allelic variation, structural divergence, and intact pathway organization^{17,18}. Moreover, the metabolic capacity of ancient gymnosperm lineages may not be completely lost at the structural level; instead, it may have undergone substantial functional attenuation, with the resulting metabolic variation potentially shaped by complex genotype-by-environment (G × E) interactions^{19,20}. Together, these considerations underscore the urgent need for integrated haplotype-resolved genomic analyses combined with targeted metabolite quantification to accurately elucidate the evolutionary and functional status of paclitaxel biosynthesis in *P. chienii*.

Here, we present a chromosome-level, haplotype-resolved genome assembly of *P. chienii* and identify a genomic architecture underlying paclitaxel biosynthesis, which we designate the Taxane Biosynthetic Archipelago (TBA). The TBA comprises multiple physically dispersed yet syntactically conserved biosynthetic gene groups and core pathway enzymes. Integrating this genomic framework with targeted metabolite quantification, transcriptomic analyses, and heterologous enzyme assays, we show that although *P. chienii* retains a largely conserved genomic architecture associated with paclitaxel biosynthesis, its terminal pathway exhibits severe functional degradation. This attenuation is characterized by reduced catalytic efficiency of PcTOT and the failure of multiple downstream enzymes to produce detectable pathway products in *Nicotiana benthamiana*. Paclitaxel is detected only at trace levels under specific conditions, whereas taxinine J and taxusin accumulate to substantial levels in specific tissues and populations. This study provides a framework for understanding the evolutionary decay and diversification of complex specialized metabolic pathways in *Taxaceae*. Furthermore, the haplotype-resolved assembly offers a valuable resource for species conservation and enables future metabolic engineering of high-value taxanes.

Results

The de novo assembly of haplotype-resolved genomes of *Pseudotsuga chienii*

To generate a high-quality genome assembly for *P. chienii*, we first conducted a genome survey. The *P. chienii* genome size was estimated

using flow cytometry and K-mer analysis. Flow cytometry suggested a size of approximately 15.64 Gb (Supplementary Fig. 2a–c and Supplementary Table 1), largely consistent with K-mer analysis (14.85 Gb) (Supplementary Fig. 2d). Furthermore, K-mer analysis showed a prominent heterozygous peak (Supplementary Fig. 2d), indicating significant genomic complexity and high heterozygosity.

To produce a high-quality assembly for *P. chienii*, we employed a multi-platform sequencing strategy, including 99 × BGI short reads, 47 × Revio HiFi long reads, and 151.42 × Hi-C sequencing for scaffolding (Supplementary Table 2). This yielded two haplotype-phased genomes (Hap1 and Hap2), both anchored to 12 pseudochromosomes, corresponding to the highly conserved haploid chromosome number ($n = 12$) of the *Taxus* species^{21,22}. The total length of the assembled pseudochromosomes is consistent with our genome size estimations, the previously published genome of *P. chienii*¹⁶. Hap1 spans 15.434 Gb (725 contigs; contig N50 60.2 Mb) and Hap2 spans 15.430 Gb (817 contigs; contig N50 55.9 Mb). (Fig. 1a and Supplementary Table 3). The Hi-C interaction heatmap exhibited a characteristic checkerboard pattern (Supplementary Figs. 3 and 4). Strong diagonal interaction patterns confirmed proper chromatin proximity and absence of aberrant contacts, which supported the accuracy of scaffold placement and the structural integrity of the assembly. The assembly length was highly consistent with the flow cytometry estimate and the K-mer analysis result, confirming comprehensive genome representation (Supplementary Table 3). Merqury analysis showed genome completeness of ~87.6% for individual haplotypes and 98.13% for merged assembly (Supplementary Table 3). High base-level consensus accuracy was also evidenced by base quality values (QV) exceeding 54 for both haplotypes (Supplementary Table 3).

De novo structural (gene model) annotation

Next, we performed high-quality genome annotation, including repeat annotation, gene prediction, and functional annotation. Comprehensive repeat annotation demonstrated that repetitive sequences were the dominant genomic component, accounting for 88.43% (Hap1) and 88.60% (Hap2) of the assembly. Among these, long terminal repeat (LTR) retrotransposons were dominant (53.65% in Hap1 and 54.05% in Hap2), with Gypsy elements and Copia elements being the most dominant (Fig. 1a and Supplementary Table 4 and 5). DNA transposons represented 21.55% (Hap1) and 21.12% (Hap2) of the genome, with Tc1/IS630/Pogo and hobo-Activator superfamilies being most abundant (Supplementary Table 4 and 5). Among LTR types, younger Gypsy elements showed greater abundance and broader genomic distribution than Copia elements, consistent with their preference for heterochromatic regions, potentially facilitating genome remodeling and regulatory diversification (Supplementary Table 4). Additionally, insertion time analysis of 33605 intact LTR retrotransposons (Hap1) revealed a unimodal expansion peak at ~21.59 MYA (Fig. 1b), which calibrated at 7.3×10^{-10} substitutions, site, and year²³, indicating singular burst of transpositional activity in the *P. chienii* genome. *P. chienii* harbors a larger genome (15.6 Gb vs. 10.2 Gb) yet fewer intact LTR insertions (31% vs. 40%) within the 8–24 MYA window in comparison to *T. mairei*⁷. This showed that genome expansion in *P. chienii* was driven by the enhanced retention of LTR that peaked several million years earlier than those in *T. mairei*, whereas the latter is distinguished by a slightly more recent burst coupled with more effective purging of young elements⁷. These findings suggested that the *P. chienii* genome exhibits extreme repetitive sequence expansion and structural complexity, likely resulting from extensive transposon amplification during evolutionary divergence, a pattern consistent with observations in other gymnosperm genomes, such as *T. grandis* and *T. mairei*^{7,24}.

Gene prediction was conducted using an integrative strategy that incorporated de novo prediction, homology-based alignment, and transcriptome evidence. A total of 36,488 and 36,360 high-confidence protein-coding genes were predicted for Hap1 and Hap2, respectively.

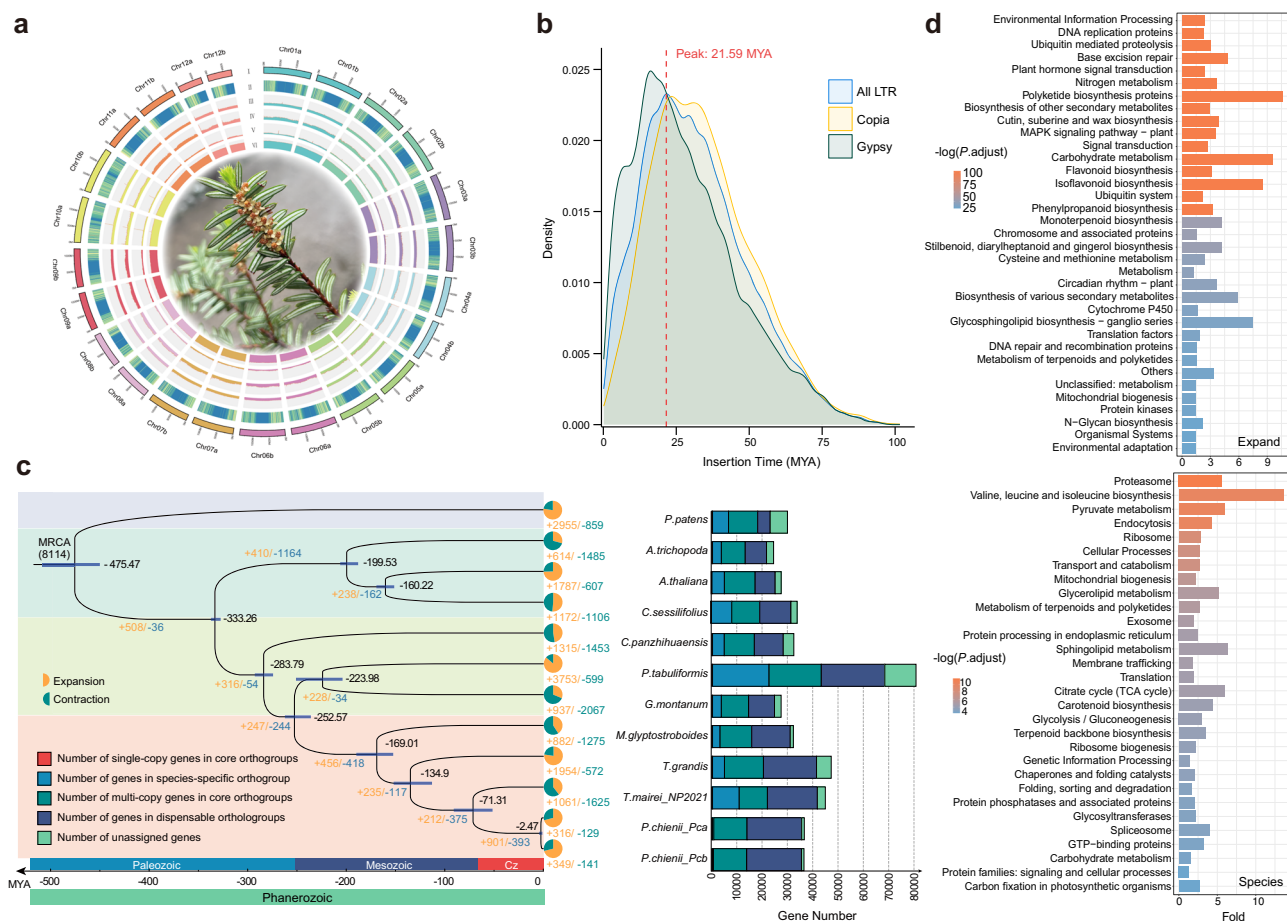


Fig. 1 | Genome characteristics and comparative genomic analysis of *Pseudotaxus chienii*. **a** Genome landscape of the 24 assembled pseudochromosomes in *P. chienii*. Track (I) displays the length of each pseudochromosome (in Mb), while tracks (II) to (VI) represent the distribution of repeat density, GC content, gene density, and the abundance of Ty3/Gypsy, Ty1/Copia, and unknown long terminal repeat (LTR) elements, respectively. All features were calculated using a 5 Mb sliding window. The central image depicts *P. chienii* during its flowering stage. **b** Expansion and diversity of LTR elements in the *P. chienii* genome. **c** Evolutionary

dynamics of gene families in *P. chienii* and representative plant species. The predicted most recent common ancestor (MRCA) is estimated to have possessed 8,146 gene families. *Pca/b*, *P. chienii* hap1/hap2. **d** KEGG pathway enrichment analysis of expanded and species-specific genes in *P. chienii*. Color intensity reflects adjusted P -value significance, and the x-axis shows enrichment fold. The P values were calculated using a two-sided hypergeometric test and adjusted for multiple comparisons using the false discovery rate (FDR). Source data are provided as a Source Data file.

BUSCO analysis based on conserved protein-coding genes revealed completeness scores of 95.0% for both haplotypes and 95.6% in combination (Fig. 1a, Supplementary Table 3 and Supplementary Fig. 5). Gene functional annotation was conducted using publicly available databases, including Nr, InterPro, Pfam, KEGG, GO, SwissProt, and eggNOG. In total, 34,332 and 34,136 genes from Hap1 and Hap2 were annotated in the Nr database, 32,976 and 32,801 in InterPro, 30,145 and 30,011 in eggNOG, 26,872 and 26,654 in Pfam, and 24,563 and 24,383 in SwissProt (Supplementary Table 6). Additionally, 25,971 (Hap1) and 25,811 (Hap2) genes were annotated with GO terms, and 9744 and 9736 genes were assigned KEGG pathway annotations. Overall, 96.13% of genes in Hap1 and 95.89% of genes in Hap2 were functionally annotated (Supplementary Table 6). These results exhibited that both haplotypes possessed high completeness and accuracy in both gene structure and function annotations.

Phylogenetic analysis revealed lineage-specific adaptation in *Pseudotaxus chienii*

To investigate *P. chienii* evolutionary trajectory and adaptive features, we conducted phylogenetic analysis based on 421 single-copy orthologs, whereas a recent study by Wang et al.¹⁶ based on a consensus

genome assembly utilized 184 single-copy genes. The results revealed a clear topology: the three *Taxaceae* species (*P. chienii*, *T. mairei*, and *T. grandis*) formed a monophyletic group that diverged from *Metasequoia glyptostroboides* (*M. glyptostroboides*) approximately 169.01 Mya (Fig. 1c). Within the *Taxaceae* clade, *P. chienii* was resolved as sister to *T. mairei* (diverging at -71.31 Mya), and this sister pair subsequently diverged from *T. grandis* at -134.9 Mya. Our estimated divergence times differ from those reported by Wang et al.¹⁶. This difference is expected and likely attributable to our distinct taxon sampling strategy and the substantially larger dataset of single-copy orthologs (421 vs. 184) utilized in our analysis. Furthermore, gene family clustering revealed 717 significantly expanded gene families in *P. chienii*, which were significantly enriched in transferase activity, polyketide biosynthesis proteins, and isoflavonoid biosynthesis (Fig. 1d and Supplementary Fig. 6). The 853 lineage-specific genes were enriched in proteasome, valine/leucine/isoleucine biosynthesis, and pyruvate metabolism. Furthermore, 3864 species-specific genes of *P. chienii* from a comparative analysis against three related conifers (*T. mairei*, *T. grandis*, and *M. glyptostroboides*) showed significant enrichment in structural constituent of ribosome, ubiquitin-protein transferase

activity, glycolysis/gluconeogenesis, ribosome, and exosome (Supplementary Fig. 7).

Conserved genomic architecture of paclitaxel biosynthetic loci in *Pseudotaxus chienii*

Previous genomic investigations in *Taxus* have revealed that a cluster of key paclitaxel biosynthetic genes is located at the terminal region of chromosome 9^{7,13} (corresponding to chromosome 10⁸ or 12⁹ in certain assemblies). To determine whether this functional genomic architecture is evolutionarily conserved in *P. chienii*, we performed a genome-wide synteny analysis involving *T. mairei*, *T. grandis* and *M. glyptostroboides*. The global genomic landscape revealed a dynamic evolutionary history between *Taxus* and *Pseudotaxus*: while chromosome 2 exhibited syntenic stability with high linearity, most other chromosomes have undergone extensive inter-chromosomal rearrangements (Fig. 2a). Under this background of chromosomal reorganization, we identified a conserved ~80.7 Mb syntenic block on the terminal region of *P. chienii* chromosome 8 of hap1 (Pca08: 1,019,626,736–1,100,348,071) that exhibits collinearity with the homologous ~80.2 Mb region in the hap2 (Pcb08: 1,015,375,398–1,095,531,480) and a strong macrosynteny with *T. mairei* chromosome 9 (Chr9: 19,994,515–91,811,351) (Fig. 2a, Supplementary Fig. 8 and 9). In contrast, the corresponding regions in *T. grandis* and *M. glyptostroboides* were fragmented, preserving only scattered orthologs of paclitaxel biosynthetic genes (e.g., *TSaH* in both, *TS* and *TI3aH* in *M. glyptostroboides*) (Fig. 2b). These results indicated that a large macro-syntenic block containing some of the core paclitaxel biosynthetic genes was specifically maintained in the genera *Taxus* and *Pseudotaxus* but was fragmented or absent in *T. grandis* and *M. glyptostroboides*.

The taxane biosynthesis pathway in *Pseudotaxus chienii* is characterized by an archipelago architecture

Leveraging the ~80.7 Mb macro-synteny block conserved across both *P. chienii* haplotypes and positionally corresponding to on *T. mairei* chromosome 9 as an anchor, a genome-wide survey of both haplotypes revealed a rich landscape of secondary metabolism, identifying 21 candidate biosynthetic gene clusters (BGCs) in Pca and 25 in Pcb (Supplementary Fig. 10, Supplementary Data 1 and 2). Consistent with patterns previously documented in gymnosperms, these clusters are exceptionally large, with an average size of approximately 4.5 Mb and approximately 65% exceeding 2 Mb in length—a scale rarely seen in angiosperms^{9,13,25}. Among them, a paclitaxel-enriched subregion on chromosome 8 contains three physically dispersed but functionally associated BGCs, designated TS1 Cluster (TS1C), Taxadiene Synthase Cluster (TSC), and T9 Cluster (T9C) within the conserved region of Pca08 (Fig. 2b). Although physically dispersed across the ~80.7 Mb block of Pca08, these three BGCs and several interspersed core paclitaxel biosynthetic genes located outside the predicted BGC boundaries are enriched in cytochrome P450s, BAHD acyltransferases, and terpene synthase enzymes, while maintaining collinearity between the two haplotypes and with the homologous region in *T. mairei*. The T9C exemplifies functional-genomic BGC definition: initially delineated at ~1.3 Mb in Pca via plantSMASH based on gene proximity, its Pcb homolog spanned ~7.7 Mb. WGCNA analysis across an integrated dataset of 124 *P. chienii* transcriptomes confirmed co-expression among metabolism-related genes across the entire 7.7 Mb Pcb region, a pattern conserved in the corresponding Pca region, thereby justifying the extension of the Pca T9C boundary to ~6.3 Mb beyond the proximity-based limit (Supplementary Fig. 11 and Supplementary Data 3).

To quantitatively validate the functional integration of this dispersed genomic archipelago, we combined chromatin architecture profiles derived from Hi-C analysis (using the Pca haplotype as a representative model) with Pearson correlation analyses of a

comprehensive 124-sample transcriptomic dataset. Hi-C profiles revealed that the entire TBA region is embedded within the transcriptionally active A compartment^{26,27}, providing a globally favorable environment for gene expression. The three gene clusters are spatially segregated into distinct Topologically Associating Domains (TADs)^{28,29}, which further subdivide into localized sub-TADs (Supplementary Fig. 12a). This hierarchical division closely aligns with functional boundaries; for example, the TAD encompassing TS1C exclusively harbors core metabolic genes, effectively demarcating the functional domain by excluding a non-metabolic gene at its left border. Although separated into independent TADs, metabolic genes from the three clusters and these dispersed core enzymes form a highly coordinated co-expression network, with the majority of gene pairs exhibiting Pearson correlation coefficients (r) ≥ 0.9 (Supplementary Fig. 12b). Collectively, these findings indicate that the functional integrity of the TBA may be supported by regulatory coordination within an active chromatin environment, rather than strictly by linear physical proximity. This archipelago-like organization, characterized by both functionally linked but spatially separated clusters and several dispersed core enzymes, contrasts with the more compact BGCs typical of microbes and angiosperms^{30,31} (Supplementary Fig. 13).

We therefore propose the term Taxane Biosynthetic Archipelago to describe this distinctive arrangement of paclitaxel biosynthetic genes in *P. chienii*, with partially conserved syntenic elements in the sister genus *Taxus*. The TBA exhibits structural plasticity in integrating conserved core modules with lineage-specific accessory elements, and is characterized by extensive serves as a dynamic region for asymmetric evolution. Each lineage has undergone extensive gene amplifications and rearrangements. For instance, the T9C (collinear with the previously defined Group 9.1 in *T. mairei*) highlights this plasticity (Fig. 2b), with the *TOT* genes varying from three copies in Pca to a single copy in Pcb, and appearing translocated near the TS1C in the *T. mairei* (Fig. 2b, c). Additionally, the Pcb haplotype may have an enhanced potential for terpene scaffold generation through the presence of nine copies of the *TS2*, an independent clade characterized by significant gene expansion and jasmonate-dependent regulatory specialization based on the classification established in *T. mairei*⁷, compared to only five in Pca (Fig. 2c). To uncover the regulatory basis of this *TS2* expansion, we analyzed promoter cis-elements across both haplotypes of *P. chienii* and *T. mairei*. A clear regulatory divergence was observed: *TS2* promoters are highly enriched in Methyl jasmonate (MeJA)-responsive elements (e.g., CGTCA-motifs, TGACG-motifs, and MYC binding sites)^{32,33} (Supplementary Fig. 14), which are largely absent in TS1. This pronounced structural divergence suggests a regulatory specialization: while TS1 likely supports baseline expression, the expanded *TS2* genes are transcriptionally primed for stress-induced responses.

Phylogenomic insights into the origin and diversification of the paclitaxel pathway

To investigate whether *P. chienii* retains the complete set of enzymes required for paclitaxel biosynthesis, we performed gene distribution analysis of 28 key genes involved in the paclitaxel biosynthetic pathway across representative land plants (Supplementary Fig. 15 and 16). *P. chienii* harbors copies of all currently known paclitaxel biosynthetic genes, indicating the retention of a largely complete genomic complement associated with paclitaxel biosynthesis. The upstream precursor biosynthesis genes geranylgeranyl diphosphate synthases (*GGPPS*), *FoTOL*, and the phenylpropanoid pathway gene phenylalanoyl-CoA ligase (*PCL*) are conserved across all surveyed land plant lineages from bryophytes to angiosperms, consistent with their inheritance from a common ancestor of embryophytes.

In contrast, the core taxane backbone-forming terpene synthases *TS1* and *TS2* are absent in bryophytes, *Gnetales* and angiosperms, but present in *Pinaceae*, most *Cupressaceae*, and *Taxaceae*. Despite their

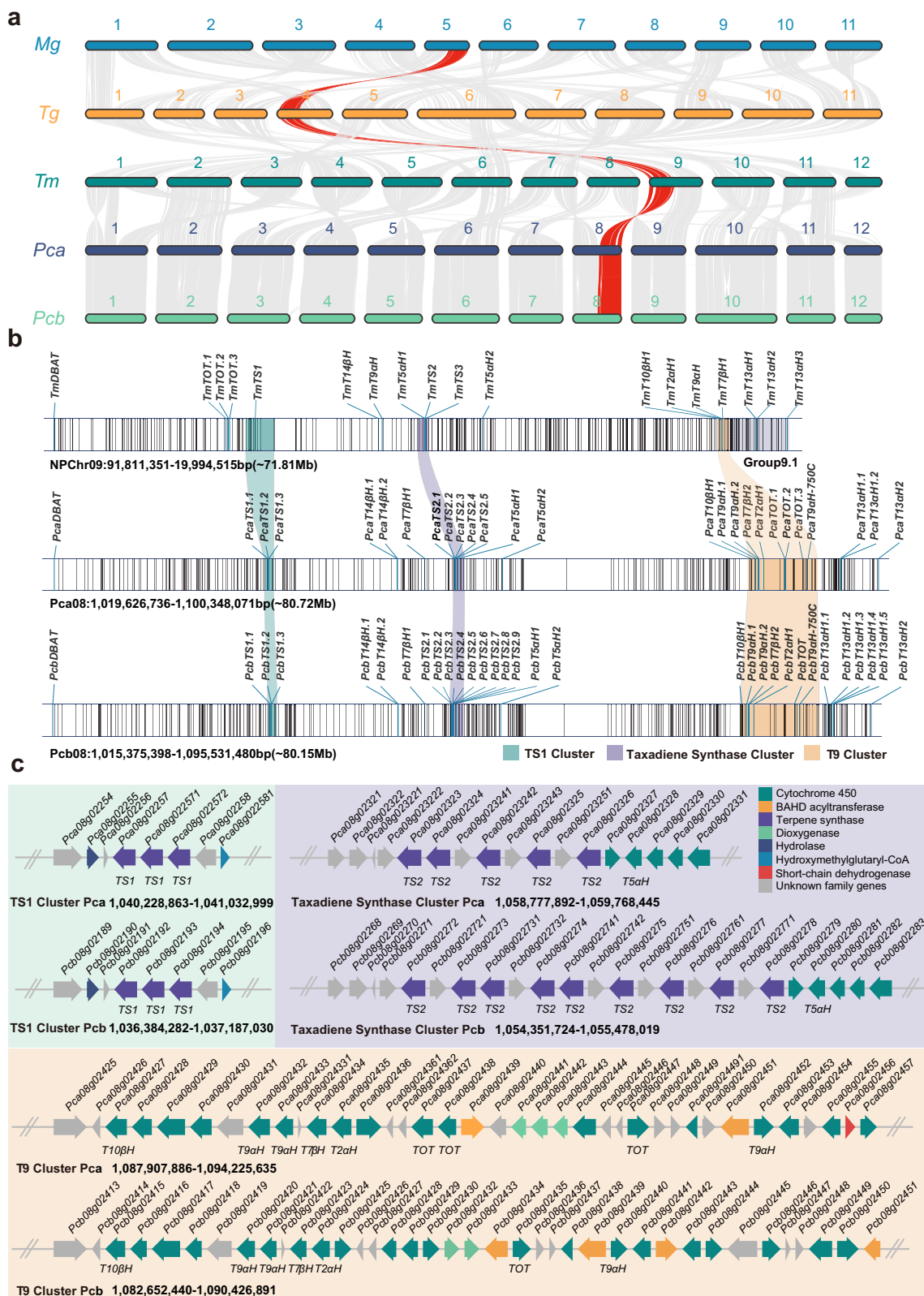


Fig. 2 | Genomic organization and structural features of Taxane Biosynthetic Archipelago (TBA) in *Pseudotsuga chienii*. **a** Macro-synteny among three *Taxaceae* species and the *Cupressaceae* outgroup *M. glyptostroboides* (Mg) highlights a conserved region containing three biosynthetic gene clusters (BGCs) and colinear upstream genes, including *TS*, *T5aH*, *T10aH* and *T13aH*. Tg, *T. grandis*; Tm, *T. mairei*; Pca/b, *P. chienii* hap1/hap2. **b** Colinear arrangement of paclitaxel biosynthetic genes

within the TBA region. Genes involved in the paclitaxel biosynthetic pathway are shown in red. **c** Organization of three paclitaxel-related BGCs on *P. chienii* chromosome Pca08. Genes are color-coded by enzymatic function, including terpene synthases, cytochrome P450s and downstream tailoring enzymes involved in taxane modification. Source data are provided as a Source Data file.

presence in these broader conifer lineages, *TS* orthologs were not detected in *Gnetum montanum*, *Cupressus gigantea*, and *T. grandis*. Specifically, the absence of these core genes in the basal *Taxaceae* species *T. grandis*, coupled with its highly fragmented syntenic region (Fig. 2a and Supplementary Fig. 8), is more consistent with secondary loss events following lineage divergence, rather than incomplete lineage sorting. This phylogenetic distribution of *TS* is potentially driven by gene duplication and subsequent functional specialization within the terpene synthase family, as demonstrated by the functional divergence of *TS* orthologs in *Taxus wallichiana* and their absence in non-Cupressales species⁵.

A second suite of downstream pathway genes involved various oxygenases, acyltransferases, and side-chain assembly enzyme (*TSaH1*, *T13aH1*, *T9aH-750C*, *T7dA*, *T9dA*, *TAX19*, *DBTNBT* and *PAM*) are present in most conifer lineages (*Pinaceae*, *Taxaceae*, and *Cupressaceae*) and absent from *Gnetales*, bryophytes, and angiosperms. The final set of pathway genes comprising additional likely functional enzymes (e.g., *T10βH1*, *T2aH*, *T7βH1*, *TOT*, *T9aH*, *T7AT*, taxane 2α-*O*-benzoyl-transferase (*TBT1*), *BAPT*) are exclusively detected in the two sequenced *Taxaceae* genomes (*T. mairei* and *P. chienii*), with 1–3 copies of each gene and zero copies in all other species surveyed.

This hierarchical distribution of these genes, spanning from ubiquitous precursors, through conifer-enriched modifications, to a suite of oxygenation, acylation and side-chain assembly enzymes, supports a step-wise evolutionary assembly of the genomic pathway complement. This indicates that the genomic architecture associated with paclitaxel biosynthesis may have been assembled in the most recent common ancestor of *Taxus* and *Pseudotsuga*. Furthermore, orthologs shared between *P. chienii* and *T. mairei* exhibited low Ka/Ks ratios (mean Ka/Ks ≈ 0.31; Supplementary Data 4), indicating sustained purifying selection on their coding sequences³⁴.

Transcriptomic and geographical and temporal targeted metabolite quantification reveal a restricted paclitaxel accumulation

To elucidate the expression pattern of paclitaxel biosynthetic genes, we performed transcriptome profiling of *P. chienii* tissues (Fig. 3a). Upstream enzymes (*GGPPS*, *TS1/2*, *FoTO1*) and the majority of pathway genes (including most hydroxylases and acyltransferases), showed higher expression in bark compared to leaves (Fig. 3a). Conversely, select genes such as *T9dA* and *T7βH1* exhibited leaf-preferential expression, suggesting that this may be a result of functional divergence in branch pathways or non-canonical taxane metabolism (Fig. 3a).

To corroborate the transcriptomic findings and further investigate the controversy surrounding the presence of paclitaxel in *P. chienii*, preliminary targeted quantification of metabolites from *P. chienii* leaf (SQS001, April 2024) was performed using a SCIEX 5500 QTRAP LC-MS/MS system and a Thermo Scientific TSQ 9000 triple-quadrupole mass spectrometer coupled with a Trace 1310 gas chromatograph for GC-MS/MS detection. A total of over 41 taxane metabolites were identified across the two detection platforms (Supplementary Fig. 17a, b), including paclitaxel and taxusin, a key intermediate in the paclitaxel biosynthetic pathway¹². Among these, taxinine J was the predominant constituent, a potential anticancer agent^{35,36}, accompanied by substantial amounts of several structurally related taxinine derivatives. Given that paclitaxel and taxusin were also detected within this complex profile, and noting the absence of the downstream intermediate baccatin III, we selected these four representative metabolites for targeted LC-MS analysis in both *P. chienii* and *T. mairei* using SCIEX 5500 LC-MS/MS platform. Importantly, baccatin III remained below the limit of detection in both leaf and bark tissues of this initial *P. chienii* individual, so only three quantifiable taxanes are included in the comparative analysis (Fig. 3b). All three compounds could be detected in *P. chienii*, with significantly higher levels in bark than leaves (Fig. 3b). Additionally, *T. mairei* leaves accumulated

substantially higher levels of paclitaxel and taxusin than any *P. chienii* tissues, whereas taxinine J accumulated to higher levels in *P. chienii* tissues than each *T. mairei* tissues. Within this initially sampled *P. chienii* individual, both paclitaxel and taxusin remained at trace amount in leaves, whereas taxinine J accumulated to substantially higher levels (Fig. 3b).

Given the trace accumulation of paclitaxel in the initial sample, we aimed to systematically determine whether this pattern was a consistent pattern across different populations of the species or a phenomenon limited to a specific region. We performed a comprehensive spatiotemporally targeted taxoid analysis on 2 distinct sampling batches from 6 independent *P. chienii* trees. The samples were taken from two geographically isolated habitats with distinct latitudes and longitudes: Sanqingshan (samples SQS001, SQS006) and Jinggangshan (samples GLM51 - GLM54) (Supplementary Fig. 18a, b). To assess temporal variation, we collected leaf and bark samples at two time points: spring 2024 (April) and winter 2026 (January), yielding a total of 18 samples (representing available tissues across seasons). All 18 spatiotemporal samples were analyzed in parallel using a high-sensitivity SCIEX Triple Quad 7500 LC-MS/MS system, while 2024 archived SQS001 samples were previously cross-validated on a SCIEX Triple Quad 5500 LC-MS/MS platform.

Across all tissues, sampling sites and time points detected, taxinine J exhibited the highest abundance among the four targeted taxanes (Supplementary Fig. 18a, b). In marked contrast, baccatin III was undetectable in all assayed samples. Trace levels of paclitaxel were exclusively found in leaves of a single *P. chienii* individual (SQS001) collected from Sanqingshan in April, 2024. Interestingly, in the Jinggangshan populations, although taxinine J maintained the highest abundance, taxusin also accumulated to a high level in selective samples, with its concentration peaking at approximately half that of taxinine J. Furthermore, following MeJA elicitation, trace paclitaxel was detected in leaves from only one out of three independent *P. chienii* branches after 48 h of treatment, whereas no paclitaxel was captured in the 96 h treatment group (Supplementary Fig. 18a), and no paclitaxel was detected any natural samples collected in 2026.

Functional characterization identifies a severe downstream bottleneck in paclitaxel biosynthesis

To uncover the biochemical basis of enzymatic function underlying the evident metabolic bottleneck and the universal absence of baccatin III in the examined samples, we systematically assessed the enzymatic capability of the paclitaxel biosynthesis pathway by transient co-expression in a heterologous *Nicotiana benthamiana* system. Guided by comparative sequence analysis, we prioritized candidate genes exhibiting lower sequence identity between *P. chienii* and *T. mairei* for functional verification (Supplementary Data 4). Specifically, we cloned *P. chienii* genes including Taxane 2α-hydroxylase (*T2αOH*), Taxane 7β-hydroxylase (*T7βOH*), taxane C-7β-*O*-acyltransferase (*T7AT*), *TBT*, and *TOT*, and transiently co-expressed them in *N. benthamiana*, with corresponding *T. mairei* genes as controls (Fig. 3c). The *P. chienii* enzymes *T2αH*, *T7βH*, *T7AT*, and *TBT* efficiently converted taxusin (1) to taxadiene hexa-acetate (2), exhibiting activities comparable to *T. mairei* orthologs (Fig. 3d, f). These heterologous expression results clearly indicate that the upstream and mid-stream steps of the paclitaxel pathway retain catalytic competence in *N. benthamiana*, whereas an obvious bottleneck arises immediately at the subsequent step. *P. chienii* *TOT* converted taxadiene hexa-acetate (2) to 1-dehydroxybaccatin IV (3) at markedly lower efficiency than *T. mairei* *TOT* (Fig. 3e, f). To preliminarily elucidate the reasons for the attenuated function of PcTOT compared to TmTOT, structural modeling and molecular docking analyses were first performed. The results revealed a slightly more relaxed catalytic conformation in PcTOT (2.7 Å/149.2°) compared to the highly optimized geometric parameters observed in TmTOT (2.4 Å/157.7°) (Supplementary Fig. 19a), despite

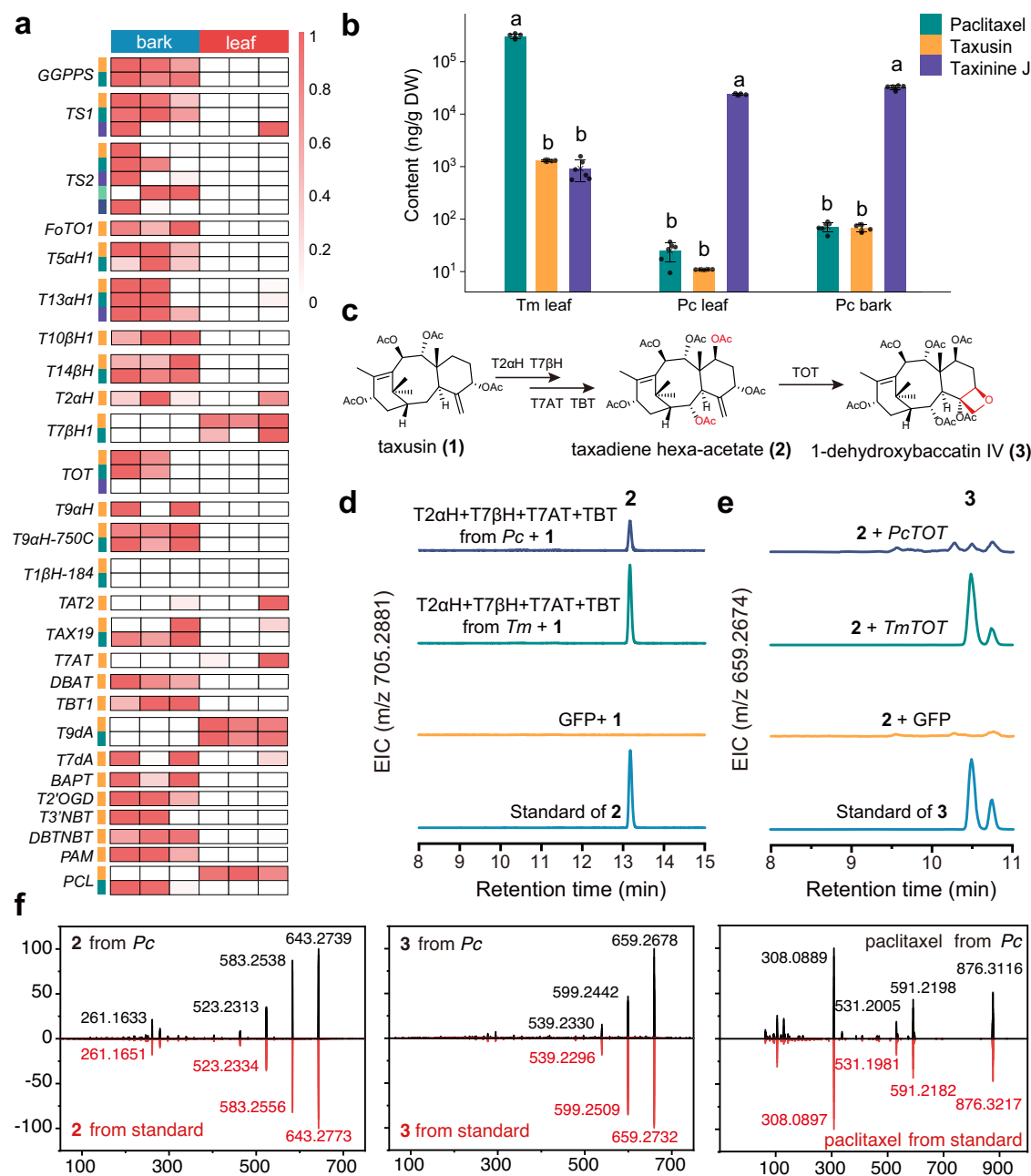


Fig. 3 | Identification of the paclitaxel biosynthetic pathway in *Pseudotaxus chienii*. **a** Heatmaps display the expression levels of corresponding biosynthetic genes in *P. chienii* leaf and *P. chienii* bark tissues. Different colors represent different gene copies in *P. chienii* hap1. **b** Targeted LC-MS profiles of key taxane metabolites in Pc leaf, Pc bark, and *T. mairei* leaf (Tm leaf). Sample sizes for Pc leaf, Pc bark, and Tm leaf were $n = 6, 6$, and 6 for paclitaxel; $n = 5, 6$, and 6 for taxusin; and $n = 4, 6$, and 6 for taxinine J, respectively. Statistical comparisons were performed on the original concentration values using one-way ANOVA followed by Fisher's least significant difference (LSD) test in OriginPro 2026 SRI (V 10.3.0.197), with $P < 0.05$ considered statistically significant. Different lowercase letters indicate a significant difference. Error bars represent means $\pm$ SD of independent biological

replicates. Note: The Y-axis is plotted using a Log10 scale to accommodate the magnitude of concentration differences between species, which visually compresses the error bars representing biological variability. **c** Schematic of T2 α H, T7 β H, T7AT, and TBT convert taxusin (**1**) to taxadiene hexa-acetate (**2**), and the schematic of TOT convert taxadiene hexa-acetate (**2**) to 1-dehydroxybaccatin IV (**3**). **d, e** EICs of expected products when the T2 α H, T7 β H, T7AT, TBT (**d**) and TOT (**e**) from *P. chienii* or *T. mairei* is expressed in *N. benthamiana*. **f** MS/MS spectra of taxadiene hexa-acetate (**2**), 1-dehydroxybaccatin IV (**3**) and paclitaxel detected in an extract of the *P. chienii* (black) sample SQS001 (2024) and its standard (red). Source data are provided as a Source Data file.

their overall structural similarity with a root-mean-square deviation (RMSD) of $1.67 \text{ \AA}$ across 2,766 backbone atoms (Supplementary Fig. 19b). Intriguingly, molecular dynamics simulations revealed distinct, region-specific flexibility patterns across these structural regions. Specifically, PcTOT exhibited a pronounced conformational mobility within the R1 segment, yet displayed a distinctly repressed flexibility, evidenced by lower calculated residue-level root-mean-

square fluctuation (RMSF) within R2 and R3 regions (Supplementary Fig. 19c). Based on CAVER analysis, these differentially rigidified regions were spatially mapped onto the residues lining the substrate access channels leading to the buried catalytic pocket (Supplementary Fig. 19d). Taken together, these complementary static and dynamic analyses suggest that the altered flexibility and subsequent channel gating dynamics likely contribute to the divergent catalytic behaviors

of the two enzymes. Moreover, despite genomic retention and high sequence conservation with *Taxus* orthologs, heterologous expression of six core downstream candidates, including *T9ox* and *BAPT*, in *N. benthamiana* failed to yield detectable pathway products, in sharp contrast to active *T. mairei* counterparts (Supplementary Fig. 20).

Discussion

The 15.43 Gb genome assembly of *P. chienii*, with its high-quality annotation, provided a valuable reference for the *Taxaceae* family and gymnosperms^{7–9,13,37}. We compared its contig N50, BUSCO completeness of protein model and repeat content with representative gymnosperms and a recently released *P. chienii* genome assembly, confirming the high reliability of our genomic dataset (Supplementary Data S5)^{7,13,16,24,37–41}. Although consensus assembly may have better linear contiguity, our haplotype-resolved assembly more effectively resolves the structural complexity of this highly heterozygous genome with over 88% repetitive sequences. Our protein annotation achieves a 95% BUSCO completeness, one of the highest among reported gymnosperms, and this high-resolution assembly facilitates the accurate reconstruction of intact metabolic pathways. The present study showed that while both *Pseudotaxus* and *Taxus* underwent significant LTR-retrotransposon expansion, their genomic structures evolved differently. The major LTR expansion in *P. chienii* (~21.59 MYA) occurred earlier than in *T. mairei*. This expansion period aligns closely with major Early Miocene climatic shifts, most notably the strengthening of the Asian monsoon⁴². Though current substitution rates based on molecular dating cannot confirm direct causality, severe environmental fluctuations likely triggered demographic bottlenecks⁴³. Such population declines could subsequently reduce the efficacy of purifying selection, permitting the extensive retention of repetitive elements⁴⁴. The larger genome of *P. chienii* was primarily attributed to the inefficient removal of ancient repetitive elements resulting from a lack of unequal recombination, rather than more frequent transposition^{45,46}. In many plants, DNA loss (through unequal recombination) balanced genome growth; however, in *P. chienii*, this imbalance between DNA gain and DNA loss has led to the long-term accumulation of repetitive sequences. This mechanism may explain how this “living fossil” maintained such an expansive genome despite a relatively ancient burst of transposon activity. As gymnosperm genomes undergo extensive expansion through the relentless proliferation of repetitive elements, the maintenance of intact biosynthetic gene clusters in linear arrays becomes increasingly difficult^{24,45}.

To address this challenge, we introduced the TBA, which illustrates a striking evolutionary adaptation: pathway coordination within these large phylogenetic lineages operates independently of physical gene proximity. Defined by functionally linked yet spatially dispersed biosynthetic gene clusters, the TBA represents a departure from the canonical model of compact BGCs³⁰, raising a key question of how functional coordination is maintained across large genomic regions. Traditional models generally invoke direct chromatin looping to account for long-range co-expression, whereas our findings reveal a distinct spatial regulatory paradigm in expanded gymnosperm genomes. Although spatially segregated into distinct TADs, the three TBA clusters maintain robust regulatory coordination within a shared active chromatin environment despite their linear separation. Instead, this coordinated expression may be supported by their localization within the transcriptionally active A compartment, coupled with unified trans-regulatory signals. This insight refines the evolutionary understanding of specialized metabolism. In large expanded genomes, the integrity of biosynthetic pathways may not require linear contiguity and may instead be maintained within broader chromatin environments. Rather, functional modules may remain topologically separated while being synchronously modulated by common global regulatory signals⁴⁷. The TBA model expanded the concept of biosynthetic gene organization, offering a complementary approach to

identify coordinated metabolic pathways in large or complex genomes where functional genes were physically dispersed, which may evade detection by conventional gene cluster search tools reliant on physical proximity. This dispersed organization may represent an adaptive feature in expansive genomes, facilitating pathway evolution amid repetitive element accumulation⁴⁶.

The haplotype-resolved assembly revealed significant genomic differences between the two haplotypes, particularly in copy number variation. For example, the Pcb haplotype contains nearly twice as many copies of the *TS2* gene as Pca. While determining the direct metabolic consequence of this specific allelic expansion is technically impractical within the blended diploid metabolome, our promoter analyses provide robust alternative evidence for its functional role. These analyses provide support for an adaptive gene dosage strategy⁴⁸: *TS2* gene expansion may enable *P. chienii* to enhance JA-responsive chemical defense upon stress, effectively balancing defensive capacity and metabolic expenditure. Furthermore, such underlying genetic variation may interact with environmental changes to shape distinct chemotypes, underscoring the need to account for G×E interactions in exploring metabolic diversification⁴⁹. Given the endangered status of *P. chienii*, these findings carry important implications for the development of more effective conservation strategies to protect its distinctive taxane biosynthetic capability. Rather than focusing exclusively on increasing population sizes, in situ conservation efforts should prioritize the protection of heterogeneous microhabitats so as to preserve G×E interactions that may contribute to chemical diversity. At the same time, ex situ germplasm collections can utilize the haplotype-resolved markers generated in this study to capture and safeguard the full range of chemotypes along with the species' underlying genetic architecture.

Resolving the historical debate surrounding paclitaxel biosynthesis in *P. chienii* requires addressing not only contradictory metabolic evidence, but also divergent genomic interpretations reported in recent studies. For instance, Wang et al.¹⁶ proposed that the pathway terminates prior to 10-deacetylbaccatin III, a conclusion that may partly reflect the absence of several corresponding pathway homologs from their consensus chromosome-scale genome assemblies. In contrast, our haplotype-resolved genome demonstrates that a largely complete genetic architecture underlying paclitaxel biosynthetic pathway, including the core TBA and other dispersed essential genes, remains structurally intact in *P. chienii*. These results further highlighted the critical importance of haplotype-aware genome assemblies for accurately reconstructing complex specialized metabolic pathways in highly heterozygous plant genomes. However, the observed discrepancies between this genomic retention and the restricted biochemical output suggest a dynamic evolutionary trajectory rather than a contradiction. While our phylogenomic analysis indicates that the genomic complement associated with the pathway was step-wise assembled and may have achieved functional integrity in the common ancestor of *Taxus* and *Pseudotaxus*, the retention of this ancient genomic architecture does not equate to modern functional execution. Following its divergence from *Taxus*, the terminal pathway in the *P. chienii* lineage appears to have undergone significant evolutionary functional degradation. The paclitaxel content in *P. chienii* leaves is over four orders of magnitude lower than that typically observed in *T. mairei*. Crucially, our findings reveal that the capacity for paclitaxel biosynthesis outside the *Taxus* genus cannot be simply reduced to a binary question of gene presence or absence. Rather, it represents a clear case of evolutionary functional degeneration. Although the ancestral archipelago architecture has been faithfully retained and the complete set of orthologous pathway genes displays low Ka/Ks ratios indicative of sustained purifying selection³⁴ and has been robustly preserved to cope with extensive genome expansion, such structural conservation does not necessarily guarantee full functional integrity. Indeed, paclitaxel could only be detected in rare instances: specifically,

in a single natural sample (SQS001) collected in April 2024 (independently detected on two LC-MS/MS platforms; the SCIEX Triple Quad 5500 LC-MS/MS spectrum is shown in Fig. 3f, and the SCIEX Triple Quad 7500 LC-MS/MS result is shown in Supplementary Fig. 18), together with only one positive case observed among all independently MeJA-induced branches in 2026 (Supplementary Fig. 18). These rare detections do not establish residual paclitaxel biosynthetic activity, although such activity cannot be completely excluded. Moreover, given the low abundance and limited reproducibility of the MeJA-associated signal, trace contamination arising during sample handling, extraction, or instrumental analysis may represent an alternative explanation for this isolated MeJA-associated detection^{50,51}. Independently, the markedly reduced catalytic activity of PcTOT and the absence of detectable activity for several downstream enzymes under the tested conditions support severe functional attenuation of the terminal pathway. Taken together, the combined biochemical and metabolomic evidence suggests that the absence of functional capacity for complete paclitaxel biosynthesis is a plausible interpretation, while weak residual activity under specific native conditions remains an unresolved possibility. The highly limited and inconsistent detection of paclitaxel may help explain the conflicting reports regarding its occurrence in *P. chienii* across previous studies^{14–16}. Resolving these alternatives and evaluating the possible contribution of G × E interactions to the observed variation in paclitaxel detection will require broader population-level and spatiotemporal sampling across multiple genotypes and environmental conditions, more extensive biological replication, stringent analytical verification, and further biochemical and *in planta* functional investigations.

To systematically investigate whether this functional deficiency extends across the downstream pathway, we characterized two core late-stage functional modules. The first module contains key oxygenases involved in skeleton modification, including T9ox (65.3% sequence identity to the *Taxus* ortholog), T9αH-750C (68.7% identity), and taxane 1β-hydroxylase (T1βH-184) (68.2% identity). The second module includes terminal acyltransferases responsible for side-chain formation and attachment, namely baccatin III:3-amino-3-phenylpropanoyl transferase (BAPT) (67.8% identity), taxoid-3'-*N*-benzoyl-transferase (T3'NBT) (73.0% identity), and 3'-*N*-debenzoyl-2'-deoxypaclitaxel-*N*-benzoyl transferase (DBTNBT) (71.9% identity). Despite genomic retention and considerable sequence conservation, none of the genes from either downstream module produced detectable pathway products in *N. benthamiana*, whereas the functional competence of the corresponding *T. mairei* orthologs remained intact (Supplementary Fig. 20).

Together with the significantly lower product formation observed for PcTOT, the absence of detectable activity across these downstream functional modules in the heterologous system indicated profound functional attenuation of the terminal paclitaxel biosynthetic pathway. Although we cannot completely exclude the possibility that these enzymes may retain weak basal activities in the presence of *P. chienii* specific co-factors or native cytochrome P450 reductases that are absent from the heterologous system, the overall heterologous *in planta* enzymatic profile, which is characterized by catalytic activities near or below the detection threshold, is consistent with our endogenous targeted metabolite analyses.

Our findings indicate that although *P. chienii* retains a largely complete genetic architecture for paclitaxel biosynthesis, this potential is not fully realized at the level of end-product accumulation. Instead, metabolic output appears to be preferentially channeled toward alternative taxane branches, particularly taxinine J. This possible metabolic rewiring is suggested by the marked accumulation of taxinine J, along with discernible levels of taxusin in specific spatiotemporal samples, and the consistent absence of downstream intermediates like baccatin III across sampled wild populations. Rather than conclusively indicating either complete evolutionary loss or retained

residual capacity, this specific taxane accumulation profile provides a biochemical basis for severe functional attenuation of the paclitaxel biosynthesis pathway. Therefore, *P. chienii* may have lost the functional capacity for complete paclitaxel biosynthesis, although a residual, trace-level capacity under specific native conditions, possibly influenced by G × E interactions, cannot be excluded and remains to be verified⁵². This phenomenon is likely rooted in the evolutionary repurposing and progressive functional degeneration of ancestral metabolic architectures⁵³.

Ultimately, the haplotype-resolved genome of the endangered *P. chienii* provides a valuable new framework for understanding how complex specialized metabolism evolves in massively expanded plant genomes. By introducing the TBA, we extend the classical concept of metabolic gene clusters, showing that within such repeat-rich gymnosperm genomes, coordinated expression of pathway-associated genes can be maintained despite their physical dispersion and may be supported by shared transcriptionally permissive chromatin environments and trans-acting regulatory signals. Importantly, the combined use of structural genomics, spatiotemporal metabolite profiling, and heterologous *in planta* functional assays helps clarify the long-standing debate over whether paclitaxel can be produced outside the *Taxus* genus. Our results point to a more subtle pattern of metabolic evolution: although the ancestral genomic architecture has been largely retained amid extreme genome expansion, pronounced functional degeneration at the later catalytic steps has constrained downstream pathway progression and is accompanied by the preferential accumulation of alternative taxane products such as taxinine J, its various structurally related derivatives (Supplementary Fig. 17), and discernible levels of taxusin across different populations. These findings thus refine our understanding of taxane diversification and offer a critical genomic foundation for both the conservation of *Taxaceae* and the engineered production of high-value taxanes.

Methods

Plant materials and sample collection

Fresh leaves and bark tissues were collected from two mature wild individual of *Pseudotsuga chienii* (accession SQS001 (male) and SQS006) located in Sanqingshan National Park (28.90087° N, 118.08942° E), Jiangxi Province, China. The sampling site was at an elevation of approximately 1,296 m above sea level. Additional fresh leaf and bark samples were collected from four mature wild individuals (accessions GLM51 - GLM54) located in Jinggangshan (26.510254° N, 114.161766° E), Jiangxi Province, China, at an elevation of approximately 1,292 m. Samples from these individuals were collected in April 2024 and January 2026. Collected tissues were immediately snap-frozen in liquid nitrogen and stored at -80 °C for subsequent analyses.

DNA and RNA extraction

High-quality genomic DNA was extracted from young leaves using a modified CTAB method. The DNA was treated with RNase A to remove RNA contaminants and further purified using AMPure XP beads (Beckman Colter, USA). DNA quantity and quality were assessed with a NanoDrop spectrophotometer, Qubit 4.0 fluorometer, agarose gel electrophoresis, and pulsed-field gel electrophoresis. Total RNA was extracted from leaf and bark tissues using the TRIzol reagent (Invitrogen, USA) and treated with RNase-free DNase I (New England Biolabs, USA) to eliminate residual genomic DNA. RNA integrity was assessed using an Agilent 2100 Bioanalyzer, and only samples with an RIN value greater than 9.0 were used for sequencing.

Library construction and sequencing

For PacBio HiFi sequencing, following an evaluation of the isolated DNA's quality, a 20 kb library was prepared with the SMRTbell Express Template Prep Kit 2.0 from Pacific Biosciences (Menlo Park, CA, USA). After passing quality checks, the resulting SMRTbell library was

sequenced on the PacBio Revio system with one 8 M SMRT Cell (Pacific Biosciences, Menlo Park, CA, USA), yielding 677.65 Gb of HiFi reads ($\sim 47.25\times$ coverage). Short-read genomic libraries were generated and sequenced on the MGI T7 platform (150 bp paired-end reads), producing 1,530.76 Gb of data ($\sim 99.2\times$ coverage). For Hi-C analysis, chromatin from formaldehyde-fixed leaf nuclei was digested using the DpnII enzyme, proximity-ligated, sheared to 300–500 bp fragments, and sequenced on the MGI T7 platform to generate 2336.39 Gb of reads ($\sim 151.42\times$ coverage). Transcriptomic libraries for RNA-seq were prepared using the VAHTS Universal V10 RNA-seq Library Prep Kit for Illumina (Vazyme, China) and sequenced on the MGI T7 platform. Iso-Seq libraries were also prepared using the SMRTbell Express Template Prep Kit 2.0 and sequenced on the PacBio Revio platform, resulting in 54.42 Gb of full-length transcriptome data. All library construction and sequencing steps adhered strictly to manufacturer protocols and quality control standards.

Genome size estimation

Genome size was estimated through flow cytometry (CytoFLEX, Beckman Colter, USA) using *Taxus mairei* as an internal reference. Three biological replicates were analyzed following propidium iodide staining. Additionally, genome size, heterozygosity, and repetitive element content were determined using K-mer (21-mer) analysis performed with GCE v1.0.2⁵⁴ on all clean MGI short-read data, evaluating K-mer depth distributions and peak features.

Genome assembly and scaffolding

HiFi reads (677.65 Gb; $\sim 47.25\times$ genome coverage) were assembled de novo using hifiasm v0.16.1⁵⁵, with Hi-C data incorporated for haplotype phasing and chromosome-scale scaffolding. Hi-C data were processed through the Juicer v1.6⁵⁶ pipeline to generate contact maps, and contigs were anchored and oriented onto pseudochromosomes using 3D-DNA v180922⁵⁷. Manual correction of misassemblies and scaffolding errors was conducted using Juicebox Assembly Tools v1.11.08⁵⁸. Additionally, the final Hi-C interaction heatmaps were generated and visualized using HapHiC v1.0.7⁵⁹. Gaps between contigs were filled using the gap-based genome filling (GBGF) method to improve assembly continuity⁶⁰. Consensus sequence polishing was performed using merfin v1.0⁶¹ based on MGI short-read data. Assembly quality was assessed using BUSCO v5.0.0⁶² (Embryophyta_odb10, 1,614 single-copy orthologs) to evaluate completeness, as well as Merqury v1.3⁶³ to calculate base-level accuracy and genome integrity, ensuring high confidence in the assembly quality.

Repeat annotation and gene prediction

Repetitive sequences were identified using RepeatModeler v2.0.5⁶⁴ to construct a species-specific transposable element (TE) library, which was combined with the Repbase database and used by RepeatMasker v4.1.5⁶⁵ for comprehensive repeat detection and masking. Gene structure prediction employed a multi-tiered strategy: (1) de novo gene prediction using Augustus v3.5.0 and BRAKER v3.0.8⁶⁶, integrating ab initio models and RNA-seq evidence; (2) homology-based prediction via GeMoMa v1.9⁶⁷ using protein sequences from *Physcomitrella patens*⁶⁰, *Amborella trichopoda* var. *SantaCruz* 75 HAP1 v2.1⁶⁸, *Oryza sativa* v7.0⁶⁸, and *Arabidopsis thaliana* araport11⁶⁸ as references; (3) transcript-based prediction involving Iso-Seq data processing followed by short-read RNA-seq assembly, both optimized with PASA v2.4.1⁶⁹. Predictions were integrated using EVIDENCEModeler v1.1.1⁶⁹, further refined by the PASA update module, and manually inspected and corrected using IGV v2.14.0⁷⁰ to produce a high-confidence gene annotation set.

Functional annotation

Functional annotation of protein-coding genes employed DIAMOND v2.0.7⁷¹ against NR⁷², Swiss-Prot⁷³, and eggNOG⁷⁴ databases (E-value $\leq$

$1e^{-5}$, alignment coverage $\geq 50\%$), complemented by domain searches using InterProScan v5.50-84.0⁷⁵ (Pfam, SMART, CDD, TIGRFAMs, ProSite). Pfam-A⁷⁶ domain confirmation was conducted using HMMER v3.4⁷⁷. Pathway annotations were generated via KEGG Automated Annotation Server (KAAS)⁷⁸ and eggNOG, with GO annotations derived from InterProScan and eggNOG ortholog groups. All functional annotations were filtered to retain those with domain coverage $\geq 50\%$ and KEGG orthology confidence ≥ 0.8 , adhering to established genome annotation protocols.

Phylogenomics and genome evolution

OrthoFinder v2.5.2⁷⁹, combined with DIAMOND, was used to identify orthologous gene families across *P. chienii* and ten representative plant genomes, including *Physcomitrella patens*, *Amborella trichopoda*, *Arabidopsis thaliana*, *Croomia sessilifolia*, *Cycas panzhihuaensis*, *Pinus tabulaeformis*, *Gnetum montanum*, *Metasequoia glyptostroboides*, *Torreya grandis*, and *T. mairei*. Among them, some gene models of *T. mairei* genome annotation were corrected. A total of 421 single-copy orthologous gene families were identified and aligned using MUSCLE v5⁸⁰. Phylogenetic trees were constructed using RAxML v8.2.12⁸¹ with the PROTGAMMAJTT model and 1,000 bootstrap replicates. Specifically, divergence times were estimated using MCMCTree (PAML v4.9j)⁸² under a relaxed molecular clock model. To calibrate the absolute evolutionary timeline, fossil calibration points obtained from the TimeTree database⁸³ were applied to specific nodes. Whole-genome duplication (WGD) events were detected using WGD1 v0.5.1⁸⁴ through collinearity and syntenic block analyses, and synonymous substitution rates (Ks) between homologous gene pairs were calculated using PAML's NG86 and YN00 methods, with Gaussian kernel density estimation applied to visualize WGD signals within and between species.

Transcriptome profiling

RNA-seq data were processed with fastp v0.23.2⁸⁵ and mapped separately onto both haplotype reference genomes (hap1 and hap2) using HISAT2 v2.2.1⁸⁶. Gene expression levels were quantified using StringTie v2.1.4⁸⁷. Functional enrichment analyses (GO and KEGG) were performed using TBtools v1.098761⁸⁸.

Identification of biosynthetic gene cluster (BGC)

The BGC were initially identified using plantSMASH v2.0.0⁸⁹ with default parameters to detect candidate clusters based on gene proximity and conserved domain signatures across both haplotypes (with same parameters) of the *P. chienii* genome. To refine these candidates and confirm functional coordination, we integrated co-expression analysis of metabolism-related genes within each cluster^{90,91}. Specifically, Weighted gene correlation network analysis (WGCNA v1.72.5) and Pearson correlation analyses were conducted in R software following the official manual⁹², to quantitatively evaluate the regulatory coherence of these clusters. To construct a highly robust, species-wide co-expression network, we utilized an integrated dataset comprising 124 *P. chienii* transcriptomes. This comprehensive dataset consists of three parts: (1) transcriptome data from a population of 108 *P. chienii* individuals⁹³ (BioProject ID: PRJNA674020), which represent diverse natural populations distinct from the individual (SQS001) sequenced in this study; (2) additional publicly available *P. chienii* transcriptomic data (BioProject ID: PRJNA630781⁹⁴, PRJEB4922⁹⁵, PRJNA726756⁹⁶, PRJNA396655⁹⁷, PRJNA498605⁹⁸); and (3) the tissue-specific transcriptomes generated in our current study (BioProject ID: PRJCA042793). Transcripts Per Kilobase Million (TPM) data for all genes across these 124 integrated samples served as input matrix. A soft-thresholding power of 9 was selected to achieve an approximate scale-free topology. The majority ($>80\%$) of the metabolism-associated genes co-assigned to the same expression module. Clusters meeting this threshold were validated as functionally coherent

BGCs, ensuring that regulatory coherence, rather than mere physical contiguity, defined the final boundaries.

Identification of orthologous and paralogous genes in the paclitaxel biosynthetic pathway

Orthologous sequences in the paclitaxel biosynthetic pathway were identified across *T. mairei*, *P. chienii*, *M. glyptostroboides*, *T. grandis*, *G. biloba*, *C. panzhihuaensis*, *Pinus tabulaeformis*, *Cupressus gigantea*, *Taxodium distichum*, *Pinus densiflora*, *Gnetum montanum*, *Physcomitrella patens*, *Amborella trichopoda* and *Arabidopsis thaliana* using the reciprocal best hit (RBH) approach⁹⁹. For each query sequence from *T. mairei*, blastp was employed to search against all protein sequences of the target species with an E-value $< 1 \times 10^{-5}$. The top-scoring hit in each target sequence was then reciprocally queried against all protein sequences of *T. mairei* using the same parameters. Pairs exhibiting reciprocal best hits and $\geq 45\%$ amino acid sequence identity were designated as orthologs. To identify paralogous genes, self-blastp searches (E-value $< 1 \times 10^{-5}$) were conducted within each species' all protein sequences. Sequences matching an ortholog with $\geq 90\%$ amino acid sequence identity and $\geq 90\%$ alignment coverage (excluding self-matches) were classified as paralogs.

Identification of chromatin compartments TADs

Cleaned paired-end Hi-C sequencing reads were processed using the HiC-Pro v3.1.0¹⁰⁰ and aligned to the *Pca* genome to construct KR-normalized contact matrices at 50 kb and 500 kb resolutions. Subsequent chromatin structure analyses were performed using the HiCExplorer v3.7.6¹⁰¹. A/B compartments were delineated using hicPCA, utilizing eigenvectors to distinguish transcriptionally active A compartments from transcriptionally repressed B compartments. Topologically Associating Domains (TADs) and their hierarchical sub-TADs boundaries were identified using the hicFindTADs tool, based on the parameters “--thresholdComparisons 0.01 --delta 0.25” and “--thresholdComparisons 0.01 --delta 0.05,” respectively. Finally, these results were visualized using PyGenomeTracks v3.9¹⁰².

Structure modeling, docking, molecular dynamics simulations, and tunnel analysis

To account for protein flexibility, the initial structures of PcTOT and TmTOT were predicted using ColabFold v1.6.1^{103,104} and subsequently expanded into conformational ensembles via AlphaFlow¹⁰⁵. Representative conformations maintaining intact active-site architectures were selected for molecular docking using AutoDock Vina v1.2.0¹⁰⁶. The docking grid was centered on the heme-containing catalytic pocket. Docked poses were evaluated based on both binding scores and catalytic geometries, specifically the distance and corresponding reaction angle between the substrate reactive atom and the catalytic oxygen. Conformations displaying favorable catalytic geometries were retained for structural comparison and downstream molecular dynamics (MD) simulations. The selected PcTOT-substrate and TmTOT-substrate complexes underwent 200-ns MD simulations using GROMACS v2022.6¹⁰⁷. The resulting trajectories were processed to remove periodic boundary conditions and structurally aligned. Root-mean-square fluctuation (RMSF) profiles were then calculated to assess residue-level flexibility. Regions showing differential local flexibility between the two enzymes were identified and mapped onto their 3D protein structures. Finally, substrate-access tunnels were computed using CAVER v3.0¹⁰⁸, with the starting point defined at the heme active center adjacent to the docked substrate.

Cloning of paclitaxel biosynthetic genes and transient expression in *N. benthamiana*

The coding sequences of *T2aH*, *T7βH*, *T7AT*, *TBT*, *TOT* were amplified from *T. mairei* and *P. chienii* cDNA by PCR. Oligonucleotide primers were purchased from Sangon Biotech, and the primer sequences are

provided in Supplementary Data 6. The PCR products were inserted into the *AgeI*- and *XhoI*- (New England Biolabs) sites of the pEAQ-HT vector. The plasmids containing the genes were transformed into *Agrobacterium tumefaciens* (strain GV3101) cells using the freeze-thaw method¹⁰⁹. The *Agrobacterium* strains were then introduced into the leaves of *N. benthamiana* plants. Each experiment was tested on leaf 6–8 (numbered by counting from the bottom) of the same *N. benthamiana* plant, as three biological replicates.

Metabolite extraction of *P. chienii* and *T. mairei* samples

For metabolite extraction, *P. chienii* and *T. mairei* tissues stored at -80°C (5–10 g) were ground in liquid nitrogen to 40–80 mesh and vacuum freeze-dried. Then, 100 mg of the resulting powder was extracted with 1.5 mL of 80% methanol. For the sample batches designated for the 7500 system, the extraction solvent was explicitly spiked with 100 ng mL⁻¹ paclitaxel-*d*₅ as a stable isotope internal standard to dynamically compensate for extraction loss and complex matrix effects. For the batches designated for the 5500 system, parallel extraction was performed without the addition of the internal standard. The mixtures were extracted by ultrasonication for 1 h at room temperature, followed by overnight incubation at 4 °C. The extracts were centrifuged at 12,000 × *g* for 10 min, and the supernatant was collected and recentrifuged. To remove any particulate matter prior to analysis, the final supernatants were filtered through a 0.22-μm nylon syringe filter and directly transferred to autosampler vials for liquid chromatography-mass spectrometry (LC-MS) injection.

Metabolite extraction of *N. benthamiana* leaves

Five days after *Agrobacterium* infiltration, *N. benthamiana* leaf tissue was collected using a leaf disc cutter (1 cm in diameter) and placed inside a 2-mL tube. Each biological replicate consisted of four disks from the same leaf, corresponding to approximately 40 mg fresh weight. For metabolite extraction, 500 μL of chilled 75% acetonitrile in water (v/v, high-performance liquid chromatography (HPLC) grade; Fisher Chemical) was added to each sample, along with one 5-mm stainless steel bead. The mixture was thoroughly homogenized using a tissue lyser at 30 Hz for 60 s to ensure complete cell lysis. Subsequently, the homogenates were centrifuged at 12,000 × *g* for 10 min at 4 °C to pellet the tissue debris. The resulting supernatants were filtered using 96-well hydrophilic PTFE filters with a pore size of 0.45 μm (Millipore) and then transferred into autosampler vials for downstream targeted metabolite detection on the SCIEX 5500 QTRAP system using the external standard method.

LC-MS analysis

Targeted metabolite detection and trace quantification of taxane compounds were strictly performed in accordance with the authoritative group standard “Quantitative detection guidelines for taxanes in *Taxus* species” (T/CI 1115-2025)¹¹⁰. The quantitative analyses were performed on a SCIEX 5500 QTRAP® triple-quadrupole/linear-ion-trap hybrid mass spectrometer equipped with a Turbo-V™ ion source and a Shimadzu LC-2030 UHPLC system. For the 5500 system, chromatographic separation was achieved on a Waters ACQUITY UPLC BEH T3 column (2.1 × 100 mm, 1.8-μm) maintained at 40 °C. Mobile phases consisted of (A) 4 mM ammonium acetate in water and (B) 4 mM ammonium acetate in methanol, delivered at 0.3 mL min⁻¹. A 32-min gradient was programmed as follows: 0–2.5 min 10 → 15% B, 2.5–26 min 15 → 90 % B, 26–26.1 min 90 → 98 % B, 26.1–28 min 98 % B, 28–28.1 min 98 → 10 % B, 28.1–30 min 10 % B, followed by 2 min of re-equilibration. Injection volume was 5 μL and samples were maintained at 8 °C in the autosampler.

To rigorously cross-validate the trace detection of paclitaxel at extreme low abundance, independent targeted analyses were also conducted on a higher-sensitivity SCIEX 7500 Triple Quad mass spectrometer coupled with an ExionLC™ system (AB SCIEX, USA),

which was equipped with a heated electrospray ionization (HESI) probe. For the 7500 system, chromatographic separation was performed on a Waters ACQUITY UPLC HSS T3 column (2.1 × 100 mm, 1.8 μm) maintained at 40 °C. The mobile phases consisted of (A) 4 mM ammonium acetate in water and (B) 4 mM ammonium acetate in methanol, delivered at a flow rate of 0.3 mL min⁻¹. A 20-min gradient program was applied as follows: 0–2.5 min, 10 → 15% B; 2.5–5 min, 15 → 40% B; 5–7 min, 40 → 70% B; 7–8 min, 70% B; 8–11 min, 70 → 80% B; 11–13 min, 80 → 100% B; 13–16 min, 100% B; 16–16.5 min, 100 → 10% B; followed by 3.5 min of re-equilibration. The injection volume was 2 μL, and the autosampler temperature was maintained at 8 °C.

The 5500 mass spectrometer was operated in positive electrospray ionization (ESI⁺) mode with multiple-reaction monitoring (MRM) using compound-specific transitions optimized via direct infusion of authentic standards. Quantification on the 5500 system was established using the external standard method. To unambiguously exclude matrix-derived false positives, enhanced product ion (EPI) scans were triggered on the 5500 QTRAP system to acquire full tandem mass spectrometry (MS/MS) fragmentation spectra, utilizing its linear-ion-trap functionality to structurally confirm the exact identity of trace paclitaxel against authentic standards.

In parallel, the ultra-sensitive 7500 system was dedicated to trace quantification of low-abundance samples utilizing the stable-isotope (paclitaxel-d₅) internal standard method in multiple reaction monitoring (MRM) mode under positive ionization, with ion transitions optimized using chemical standards. Source parameters for the 5500 system were: curtain gas 25 psi, ion-spray voltage 5.5 kV, temperature 500 °C, nebulizer gas (GS1) 50 psi, and heater gas (GS2) 50 psi. Source parameters for the 7500 system were: curtain gas 40 psi, ion-spray voltage 2.5 kV, temperature 400 °C, nebulizer gas (GS1) 35 psi, and heater gas (GS2) 60 psi. Compound-dependent parameters (including declustering potential, collision energy, and cell-exit potential) were automatically optimized with Analyst[®] 1.7 and SCIEX OS software.

Calibration curves (0.5–1000 ng mL⁻¹) of taxinine J (CAS: 18457-46-0), taxusin (CAS: 19605-80-2), Baccatin III (CAS: 27548-93-2) and paclitaxel (CAS: 33069-62-4) standards, as well as other target metabolites analyzed in *N. benthamiana* leaves, were prepared by serial dilution of authentic standards in 80% methanol to match the sample solvent matrix. For the 7500 system calibration, the standard solutions were consistency spiked with 20 ng mL⁻¹ paclitaxel-d₅ to match the internal standard concentration in the samples. Quality-control pools (low, mid, high) were injected every 10 samples to actively monitor system stability and verify the absence of background interference; intra-day RSDs were <14.8 %. Data acquisition and quantification were performed with Analyst 1.7 and MultiQuant 3.0 softwares for the 5500 system, applying external standard calibration. For the 7500 system, SCIEX OS software was applied for taxinine J, taxusin, Baccatin III and paclitaxel identification and peak integration, utilizing internal-standard normalization with a 1/x² weighting factor.

GC-MS analysis of taxadiene and Taxa-4(20),11-dien-5α-ol

For metabolite extraction, the plant tissues ground in liquid nitrogen were accurately weighed (approximately 20 mg) into a 1.5-mL centrifuge tube, followed by the addition of 1 mL of ethyl acetate. The mixture was vortexed for 3 min and subjected to ultrasonication at 25 °C and 200 W for 30 min. The sample was then centrifuged at 12,000 × g for 5 min. A 1-mL aliquot of the supernatant was precisely transferred to a separate tube containing 100 mg of anhydrous magnesium sulfate (MgSO₄) to remove residual moisture and polar matrix interferences. After vortexing for 3 min and centrifuging at 12,000 × g for 5 min, the final supernatant was filtered through a 0.22-μm organic nylon syringe filter and transferred directly into an autosampler vial for gas chromatography-mass spectrometry (GC-MS) analysis.

Quantitative analysis was performed on a Thermo Scientific TSQ 9000 triple-quadrupole mass spectrometer coupled with a Trace 1310 gas chromatograph system. Chromatographic separation was achieved on a DB-5MS UI capillary column (30 m × 0.25 mm, 0.25 μm film thickness; Agilent Technologies). The injection volume was 1 μL with a split ratio of 20:1. High-purity helium (99.999%) was used as the carrier gas at a constant flow rate of 1.2 mL min⁻¹, and the injector temperature was maintained at 270 °C. The oven temperature program was set as follows: the initial temperature was held at 100 °C for 1 min, then increased to 175 °C at a rate of 15 °C/min, subsequently raised to 220 °C at a rate of 4 °C/min and kept constant for 2 min, and finally raised to 290 °C at a rate of 20 °C/min and held constant for 5 min. The MS transfer line and electron ionization (EI) source temperatures were set at 250 °C and 300 °C, respectively. The mass spectrometer was operated in EI mode at 70 eV, utilizing selected ion monitoring (SIM) for the highly sensitive detection of target compounds.

Calibration curves for taxadiene (CAS: 109808-37-9) and taxa-4(20),11-dien-5α-ol (CAS: 148381-56-0) were prepared by serial dilution of authentic standards in a hexane:ethyl acetate mixture (4:1, v/v) to yield concentrations of 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, and 100 μg mL⁻¹. The standard solutions were injected sequentially, and data acquisition and quantification were performed using the Chromeleon software suite. Calibration curves were constructed by plotting peak areas against the corresponding analyte concentrations to establish linear regression equations, which were subsequently used to calculate the absolute contents of the two target taxanes in the samples.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All sequencing data, including genomic sequencing, Hi-C, RNA-seq, and Revio HiFi reads, have been deposited in the NGDC Sequence Read Archive (SRA) under BioProject [PRJCA042793](https://doi.org/10.1038/s41467-026-77115-w). The de novo genome assembly and annotation have been deposited in Figshare (<https://doi.org/10.6084/m9.figshare.30340690>). The mass spectrometry-based metabolomics data, including LC-MS/MS and GC-MS data, have been deposited in OMIX under accession number OMIX019547 (<https://ngdc.cncb.ac.cn/omix/release/OMIX019547>). Source data are provided with this paper.

Code availability

The custom scripts and pipelines used for the genome assembly, haplotype phasing, and evolutionary analyses of paclitaxel biosynthesis in this study are publicly available in the GitHub repository (<https://github.com/wanghuan766/Pseudotaxus-chienii-genome-project>) and also archived on Zenodo (<https://doi.org/10.5281/zenodo.21839213>)¹¹¹.

References

1. Fu, L., Li, N. & Mill, R. Taxaceae. *Flora China* **4**, 89–96 (1999).
2. Robertshaw, B., Grimshaw, J. & Christian, T. *Pseudotaxus chienii*, Trees and Shrubs Online <https://treesandshrubsonline.org/> (2020).
3. Dörken, V. M. & Nimsch, H. *Morphology, anatomy and systematics of Pseudotaxus (Taxaceae, Coniferales) – Taxus-like, but not a Taxus* (2022).
4. Coughlan, P., Carolan, J. C., Hook, I. L., Kilmartin, L. & Hodkinson, T. R. Phylogenetics of *Taxus* using the internal transcribed spacers of nuclear ribosomal DNA and plastid trnL-F regions. *Horticulturae* **6**, 19 (2020).

5. Xie, X., Zhai, Y., Cheng, H., Wei, W. H. & Ren, M. From *Taxus* to paclitaxel: opportunities and challenges for urban agriculture to promote human health. *Plant Physiol. Biochem.* **220**, 109502 (2025).
6. Fernie, A. R., Liu, F. & Zhang, Y. Post-genomic illumination of paclitaxel biosynthesis. *Nat. Plants* **10**, 1875–1885 (2024).
7. Xiong, X. et al. The *Taxus* genome provides insights into paclitaxel biosynthesis. *Nat. Plants* **7**, 1026–1036 (2021).
8. Cheng, J. et al. Chromosome-level genome of Himalayan yew provides insights into the origin and evolution of the paclitaxel biosynthetic pathway. *Mol. Plant* **14**, 1199–1209 (2021).
9. Song, C. et al. *Taxus yunnanensis* genome offers insights into gymnosperm phylogeny and taxol production. *Commun. Biol.* **4**, 1203 (2021).
10. Liang, F. et al. Elucidation of the final steps in Taxol biosynthesis and its biotechnological production. *Nat. Synth.* **4**, 1212–1222 (2025).
11. Jiang, B. et al. Characterization and heterologous reconstitution of *Taxus* biosynthetic enzymes leading to baccatin III. *Science* **383**, 622–629 (2024).
12. McClune, C. J. et al. Discovery of FoTO1 and Taxol genes enables biosynthesis of baccatin III. *Nature* **643**, 582–592 (2025).
13. Li, Z. et al. Phased high-quality genome of the gymnosperm Himalayan Yew assists in paclitaxel pathway exploration. *Giga-science* **14** (2025).
14. Heinig, U., Scholz, S. & Jennewein, S. Getting to the bottom of Taxol biosynthesis by fungi. *Fungal Divers* **60**, 161–170 (2013).
15. Yu, C. et al. Omic analysis of the endangered *Taxaceae* species *Pseudotaxus chienii* revealed the differences in taxol biosynthesis pathway between *Pseudotaxus* and *Taxus yunnanensis* trees. *BMC Plant Biol.* **21**, 104 (2021).
16. Wang, M. et al. Analysis genome of *Pseudotaxus chienii* reveals insights into the origin and evolution of taxane biosynthesis. *Nat. Commun.* **17**, 1088 (2025).
17. He, H. et al. Haplotype-resolved genome of a citronella provides insights into the evolution of citronellol biogenesis pathway. *Hortic. Res.* **13**, uhaf287 (2026).
18. Han, D. et al. Pan-genome assembly of vine tea (*Nekemias grossedentata*) reveals structural variation in its dihydromyricetin biosynthesis diversity. *Hortic. Res.* **13**, uhaf307 (2026).
19. Han, X. et al. Genetic and transcriptome analyses of the effect of genotype-by-environment interactions on *Brassica napus* seed oil content. *Plant Cell* **37**, koaf062 (2025).
20. Meteignier, L. V. et al. Harnessing the spatial and transcriptional regulation of monoterpenoid indole alkaloid metabolism in *Alstonia scholaris* leads to the identification of broad geissoschizine cyclase activities. *Plant Physiol. Biochem.* **219**, 109363 (2025).
21. Tomasino, M. P., Gennaro, A., Simeone, M. C., Schirone, B. & Ceoloni, C. New insights into the *Taxus baccata* L. karyotype based on conventional and molecular cytogenetic analyses. *Caryologia* **70**, 248–257 (2017).
22. He, Z., Luo, X., Lei, Y. & Zhang, W. Five species of *Taxus* karyotype based on oligo-FISH for 5S rDNA and (AG)₃T₃. *Genes* **13**, 12 (2022).
23. De La Torre, A. R., Li, Z., Van de Peer, Y. & Ingvarsson, P. K. Contrasting rates of molecular evolution and patterns of selection among gymnosperms and flowering plants. *Mol. Biol. Evol.* **34**, 1363–1377 (2017).
24. Niu, S. et al. The Chinese pine genome and methylome unveil key features of conifer evolution. *Cell* **185**, 204–217.e14 (2022).
25. Forman, V. et al. A gene cluster in *Ginkgo biloba* encodes unique multifunctional cytochrome P450s that initiate ginkgolide biosynthesis. *Nat. Commun.* **13**, 5143 (2022).
26. Dong, P. et al. 3D chromatin architecture of large plant genomes determined by local A/B compartments. *Mol. Plant* **10**, 1497–1509 (2017).
27. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
28. Liu, C., Cheng, Y. J., Wang, J. W. & Weigel, D. Prominent topologically associated domains differentiate global chromatin packing in rice from Arabidopsis. *Nat. Plants* **3**, 742–748 (2017).
29. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
30. Cawood, G. L. & Ton, J. Decoding resilience: ecology, regulation, and evolution of biosynthetic gene clusters. *Trends Plant Sci.* **30**, 185–198 (2025).
31. Seshadri, K. et al. Synthetic biology in natural product biosynthesis. *Chem. Rev.* **125**, 3814–3931 (2025).
32. Dombrecht, B. et al. MYC2 differentially modulates diverse jasmonate-dependent functions in Arabidopsis. *Plant Cell* **19**, 2225–2245 (2007).
33. Lescot, M. et al. PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Res.* **30**, 325–327 (2002).
34. Hurst, L. D. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet.* **18**, 486 (2002).
35. Kobayashi, J. & Shigemori, H. Bioactive taxoids from the Japanese yew *Taxus cuspidata*. *Med. Res. Rev.* **22**, 305–328 (2002).
36. Hosoyama, H., Shigemori, H., Tomida, A., Tsuruo, T. & Kobayashi, J. Modulation of multidrug resistance in tumor cells by taxinine derivatives. *Bioorg. Med. Chem. Lett.* **9**, 389–394 (1999).
37. Lou, H. et al. The *Torreya grandis* genome illuminates the origin and evolution of gymnosperm-specific sciadonic acid biosynthesis. *Nat. Commun.* **14**, 1315 (2023).
38. Liu, Y. et al. The *Cycas* genome and the early evolution of seed plants. *Nat. Plants* **8**, 389–401 (2022).
39. Liu, H. et al. The nearly complete genome of *Ginkgo biloba* illuminates gymnosperm evolution. *Nat. Plants* **7**, 748–756 (2021).
40. Neale, D. B. et al. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biol.* **15**, R59 (2014).
41. Fu, F. et al. The *Metasequoia* genome and evolutionary relationships among redwoods. *Plant Commun.* **4**, 100643 (2023).
42. Guo, Z. T. et al. Onset of asian desertification by 22 Myr ago inferred from loess deposits in China. *Nature* **416**, 159–163 (2002).
43. Casacuberta, E. & Gonzalez, J. The impact of transposable elements in environmental adaptation. *Mol. Ecol.* **22**, 1503–1517 (2013).
44. Lynch, M. & Conery, J. S. The origins of genome complexity. *Science* **302**, 1401–1404 (2003).
45. Nystedt, B. et al. The Norway spruce genome sequence and conifer genome evolution. *Nature* **497**, 579–584 (2013).
46. Wan, T. et al. Evolution of complex genome architecture in gymnosperms. *Gigascience* **11**, giac078 (2022).
47. Nutzmann, H. W. et al. Active and repressed biosynthetic gene clusters have spatially distinct chromosome states. *Proc. Natl. Acad. Sci. USA* **117**, 13800–13809 (2020).
48. Lye, Z. N. & Purugganan, M. D. Copy number variation in domestication. *Trends Plant Sci.* **24**, 352–365 (2019).
49. Zust, T. & Agrawal, A. A. Trade-offs between plant growth and defense against insect herbivory: an emerging mechanistic synthesis. *Annu. Rev. Plant Biol.* **68**, 513–534 (2017).
50. European Commission. Quality control procedures for pesticide residues analysis (Document No. SANCO/10476/2003). Guidelines for Residues Monitoring in the European Union (2004).
51. Codex Alimentarius Commission. CAC/GL 40–1993: Guidelines on Good Laboratory Practice in Pesticide Residue Analysis (Rome, 2010).

52. Scherlach, K. & Hertweck, C. Mining and unearthing hidden bio-synthetic potential. *Nat. Commun.* **12**, 3864 (2021).
53. Moghe, G. D. & Last, R. L. Something old, something new: conserved enzymes and the evolution of novelty in plant specialized metabolism. *Plant Physiol.* **169**, 1512–1523 (2015).
54. Liu, B. et al. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. Preprint at <https://arxiv.org/abs/arXiv:1308.2012> (2013).
55. Cheng, H. et al. Haplotype-resolved assembly of diploid genomes without parental data. *Nat. Biotechnol.* **40**, 1332–1335 (2022).
56. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
57. Dudchenko, O. et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
58. Durand, N. C. et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
59. Zeng, X. et al. Chromosome-level scaffolding of haplotype-resolved assemblies using Hi-C data without reference genomes. *Nat. Plants* **10**, 1184–1200 (2024).
60. Bi, G. et al. Near telomere-to-telomere genome of the model plant *Physcomitrium patens*. *Nat. Plants* **10**, 327–343 (2024).
61. Formenti, G. et al. Merfin: improved variant filtering, assembly evaluation and polishing via k-mer validation. *Nat. Methods* **19**, 696–704 (2022).
62. Manni, M., Berkeley, M. R., Seppey, M., Simao, F. A. & Zdobnov, E. M. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).
63. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
64. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451–9457 (2020).
65. Chen, N. et al. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **Chapter 4**, Unit 4 10 (2004).
66. Gabriel, L. et al. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res* **34**, 769–777 (2024).
67. Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S. O. & Grau, J. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinform* **19**, 189 (2018).
68. Goodstein, D. M. et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, D1178–D1186 (2012).
69. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVIDENCEModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
70. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
71. Buchfink, B., Reuter, K. & Drost, H. G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
72. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, D61–D65 (2007).
73. Boeckmann, B. et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31**, 365–370 (2003).
74. Huerta-Cepas, J. et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* **47**, D309–D314 (2019).
75. Mitchell, A. L. et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* **47**, D351–D360 (2019).
76. Mistry, J. et al. Pfam: The protein families database in 2021. *Nucleic Acids Res* **49**, D412–D419 (2021).
77. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
78. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* **35**, W182–W185 (2007).
79. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
80. Edgar, R. C. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nat. Commun.* **13**, 6968 (2022).
81. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinform* **30**, 1312–1213 (2014).
82. dos Reis, M. & Yang, Z. Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. *Mol. Biol. Evol.* **28**, 2161–2172 (2011).
83. Kumar, S. et al. TimeTree 5: an expanded resource for species divergence times. *Mol. Biol. Evol.* **39** (2022).
84. Sun, P. et al. WGDl: A user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. *Mol. Plant* **15**, 1841–1851 (2022).
85. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinform* **34**, i884–i890 (2018).
86. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
87. Shumate, A., Wong, B., Perte, G. & Perte, M. Improved transcriptome assembly using a hybrid of long and short reads with StringTie. *PLoS Comput. Biol.* **18**, e1009730 (2022).
88. Chen, C. et al. TBtools-II: A “one for all, all for one” bioinformatics platform for biological big-data mining. *Mol. Plant* **16**, 1733–1742 (2023).
89. Del Pup, E. et al. plantiSMASH 2.0: improvements to detection, annotation, and prioritization of plant biosynthetic gene clusters. *J. Mol. Biol.* **438**, 169798 (2026).
90. Wisecaver, J. H. et al. A Global Coexpression Network Approach for Connecting Genes to Specialized Metabolic Pathways in Plants. *Plant Cell* **29**, 944–959 (2017).
91. Priego-Cubero, S. et al. Subfunctionalization and epigenetic regulation of a biosynthetic gene cluster in Solanaceae. *Proc. Natl. Acad. Sci. USA* **122**, e2420164122 (2025).
92. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform* **9**, 1–13 (2008).
93. Liu, L., Wang, Z., Su, Y. & Wang, T. Population transcriptomic sequencing reveals allopatric divergence and local adaptation in *Pseudotaxus chienii* (Taxaceae). *BMC Genom.* **22**, 388 (2021).
94. Liu, L., Wang, Z., Su, Y. & Wang, T. Characterization and Analysis of the Full-Length Transcriptomes of Multiple Organs in *Pseudotaxus chienii* (W.C.Cheng) W.C.Cheng. *Int. J. Mol. Sci.* **21**, (2020).
95. Carpenter, E. J. et al. Access to RNA-sequencing data from 1,173 plant species: the 1000 Plant transcriptomes initiative (1KP). *Gigascience* **8**, giz126 (2019).
96. Stull, G. W. et al. Gene duplications and phylogenomic conflict underlie major pulses of phenotypic evolution in gymnosperms. *Nat. Plants* **7**, 1015–1025 (2021).
97. Ran, J. H., Shen, T. T., Wang, M. M. & Wang, X. Q. Phylogenomics resolves the deep phylogeny of seed plants and indicates partial

- convergent or homoplastic evolution between Gnetales and angiosperms. *Proc. Biol. Sci.* **285**, 20181012 (2018).
98. Yu, C. et al. Comparative metabolomic analyses revealed the differential accumulation of taxoids, flavonoids and hormones among six *Taxaceae* trees. *Sci. Hortic.* **285**, 110196 (2021).
 99. Nestor, B. J., Bayer, P. E., Fernandez, C. G. T., Edwards, D. & Finnegan, P. M. Approaches to increase the validity of gene family identification using manual homology search tools. *Genetica* **151**, 325–338 (2023).
 100. Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
 101. Ramirez, F. et al. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat. Commun.* **9**, 189 (2018).
 102. Lopez-Delisle, L. et al. pyGenomeTracks: reproducible plots for multivariate genomic datasets. *Bioinform* **37**, 422–423 (2021).
 103. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
 104. Mirdita, M. et al. ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
 105. Jing, B., Berger, B. & Jaakkola, T. AlphaFold meets flow matching for generating protein ensembles. Preprint at <https://arxiv.org/abs/arXiv:2402.04845> (2024).
 106. Eberhardt, J., Santos-Martins, D., Tillack, A. F. & Forli, S. AutoDock Vina 1.2.0: new docking methods, expanded force field, and python bindings. *J. Chem. Inf. Model.* **61**, 3891–3898 (2021).
 107. Pall, S. et al. Heterogeneous parallelization and acceleration of molecular dynamics simulations in GROMACS. *J. Chem. Phys.* **153**, 134110 (2020).
 108. Chovanova, E. et al. CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. *PLoS Comput. Biol.* **8**, e1002708 (2012).
 109. Holsters, M. et al. Transfection and transformation of *Agrobacterium tumefaciens*. *Mol. Gen. Genet.* **163**, 181–187 (1978).
 110. China International Association for Promotion of Science and Technology. Quantitative detection guidelines for taxanes in *Taxus* species. T/CI 1115—2025 (China International Association for Promotion of Science and Technology, Beijing, 2025).
 111. Wang H. et al. Pseudotaxus-chienii-genome-project: Pseudotaxus chienii genome project v1.0.0". *Zenodo* <https://doi.org/10.5281/zenodo.21839213> (2026).

Acknowledgements

We thank Lingling Chen and Bao Nie at the Agricultural Genomics Institute at Shenzhen for their assistance with MS analysis and flow cytometry experiments, respectively.

Author contributions

J.B.Y. initiated and supervised the project. H.W. designed and implemented the genome assembly, evaluation, and data analysis strategies. H.D.C. performed the comparative genomics analysis, and Y.P.S. performed the experiments. M.F.S., Y.Z., W.Y.X., R.D., Y.G. L., Y.C. X., S.C.L., B.J., S.X.B., and Y.C.P. assisted with metabolic assays and advised on the project. L.M.G. provided plant materials. H.W. analyzed data and wrote

the manuscript. J.B.Y., J.M.L. and H.W. revised the manuscript. All the authors discussed and commented on the manuscript.

Funding

This research was financially supported by the National Key R&D Program of China projects (2023YFA0915800 and 2025YFA0921504 to J.B.Y.); the National Natural Science Foundation of China (Grant No. 32425011 to J.B.Y. and 32502746 to J.M.L.); the Agricultural Science and Technology Innovation Program (ASTIP, No. CAAS-CSIAF-202302) to J.B.Y.; the Beijing Life Science Academy (BLSA) (2024200CA0020) to J.B.Y.; and the New Cornerstone Science Foundation through the XPLOER PRIZE to J.B.Y.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-026-77115-w>.

Correspondence and requests for materials should be addressed to Jiaming Li or Jianbin Yan.

Peer review information *Nature Communications* thanks Benoit St-Pierre, who co-reviewed with Mickael DurandWen Wang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.